

566.43181X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): IWAMURA, et al.
Serial No.: Not assigned
Filed: October 2, 2003
Title: MULTI-SITE REMOTE-COPY SYSTEM
Group: Not assigned

LETTER CLAIMING RIGHT OF PRIORITY

Mail Stop Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

October 2, 2003


Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Application No.(s) 2003-207004 filed August 11, 2003.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Carl I. Brundidge
Registration No. 29,621

CIB/amr
Attachment
(703) 312-6600

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 8月11日
Date of Application:

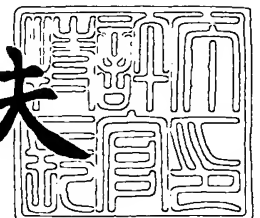
出願番号 特願2003-207004
Application Number:
[ST. 10/C] : [JP 2003-207004]

出願人 株式会社日立製作所
Applicant(s):

2003年 9月26日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2003-3079433

【書類名】 特許願

【整理番号】 K03009441A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 12/00

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 岩村 卓成

【発明者】

 【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所システム開発研究所内

 【氏名】 大枝 高

【発明者】

 【住所又は居所】 神奈川県小田原市中里 3 3 2 番地 2 号 株式会社日立製作所 R A I D システム事業部内

 【氏名】 佐藤 孝夫

【特許出願人】

 【識別番号】 000005108

 【氏名又は名称】 株式会社 日立製作所

【代理人】

 【識別番号】 100075096

 【弁理士】

 【氏名又は名称】 作田 康夫

【手数料の表示】

 【予納台帳番号】 013088

 【納付金額】 21,000円

【提出物件の目録】

 【物件名】 明細書 1

 【物件名】 図面 1

【物件名】	要約書	1
【プルーフの要否】	要	

【書類名】 明細書

【発明の名称】 複数のサイトにリモートコピーを行うシステム

【特許請求の範囲】

【請求項 1】

計算機及び記憶装置を有するシステムにおいて、
前記計算機が前記記憶装置へデータを二重化して複数の記憶領域へ書き込み、
前記記憶装置は、前記二重化して書き込まれた前記複数の記憶領域のうち第一の記憶領域へのデータ更新の内容を該記憶装置と接続される第二の記憶装置へ、該記憶領域への前記計算機の前記データ更新の要求が完了する前に転送し、
前記記憶装置は、前記二重化して書き込まれた前記複数の記憶領域のうち第二の記憶領域への前記データ更新の内容を該記憶装置と接続される第三の記憶装置へ、該記憶領域への前記計算機の前記データ更新の要求の完了後に転送することを特徴とする計算機システム。

【請求項 2】

前記記憶装置は、前記第二の記憶装置との間の接続に異常が発生した場合は、前記計算機の前記第一の記憶領域及び前記第二の記憶領域のデータの更新要求を受け付けないことを特徴とする請求項 1 記載の計算機システム。

【請求項 3】

前記記憶装置は、前記第二の記憶装置からデータを受信して前記第一の記憶領域に格納されていたデータを再構成することを特徴とする請求項 2 記載の計算機システム。

【請求項 4】

前記記憶装置は、前記第三の記憶装置からデータを受信して前記第一の記憶領域に格納されていたデータを再構成することを特徴とする請求項 2 記載の計算機システム。

【請求項 5】

第一のサイト、第二のサイト及び第三のサイトを有するシステムにおいてデータを複製する方法であって、

各サイトは計算機及び記憶装置を有し、

前記第一のサイトでデータを二重化して第一及び第二の記憶領域へ保存し、
前記第一の記憶領域の更新データを前記第二のサイトへ同期リモートコピーで転送し、

前記第二の記憶領域の更新データを前記第三のサイトへ非同期リモートコピーで転送することを特徴とするデータの複製方法。

【請求項 6】

前記第一のサイトに障害が発生した場合に、
前記第二のサイトに含まれる計算機で前記第一のサイトに含まれる計算機で行われていた処理を継続し、

前記第二のサイトが有する記憶装置の記憶領域の更新データを前記第三のサイトへ転送することを特徴とする請求項 5 記載のデータ複製方法。

【請求項 7】

前記第一のサイトが復旧した場合、前記第二のサイトの計算機で実行されていた処理を前記第一のサイトの計算機が継続し、前記第二のサイトの記憶装置に格納されていたデータを前記第一のサイトが有する記憶装置へ転送し、前記第一のサイトは、前記二重化、前記同期リモートコピー及び前記非同期リモートコピーの処理を再開することを特徴とする請求項 6 記載のデータ複製方法。

【請求項 8】

前記第一のサイトが復旧した場合、前記第二のサイトの計算機で実行されていた処理を前記第一のサイトの計算機が継続し、前記第三のサイトの記憶装置に格納されていたデータを前記第一のサイトが有する記憶装置へ転送し、前記第一のサイトは、前記二重化、前記同期リモートコピー及び前記非同期リモートコピーの処理を再開することを特徴とする請求項 6 記載のデータ複製方法。

【請求項 9】

前記第一のサイトに障害が発生した場合に、前記第三のサイトに含まれる計算機で前記第一のサイトに含まれる計算機で行われていた処理を継続し、

前記第二のサイトが有する記憶装置に格納されるデータを前記第三のサイトへ転送して前記第二及び第三のサイトが有する記憶装置のデータの内容を一致させ、前記第三のサイトの記憶装置へのデータ更新の内容を前記第二のサイトの記憶

装置へ転送することを特徴とする請求項 5 記載のデータ複製方法。

【請求項 10】

前記第一のサイトに障害が発生した場合に、前記第三のサイトに含まれる計算機で前記第一のサイトに含まれる計算機で行われていた処理を継続し、

前記第三のサイトの記憶装置へのデータ更新の内容を前記第二のサイトの記憶装置へ転送することを特徴とする請求項 5 記載のデータ複製方法。

【請求項 11】

前記第一のサイトが復旧した場合、前記第三のサイトの計算機で実行されていた処理を前記第一のサイトの計算機が継続し、前記第三のサイトの記憶装置に格納されていたデータを前記第一のサイトが有する記憶装置へ転送し、前記第一のサイトは、前記二重化、前記同期リモートコピー及び前記非同期リモートコピーの処理を再開することを特徴とする請求項 10 記載のデータ複製方法。

【請求項 12】

前記第一のサイトが復旧した場合、前記第三のサイトの計算機で実行されていた処理を前記第一のサイトの計算機が継続し、前記第二のサイトの記憶装置に格納されていたデータを前記第一のサイトが有する記憶装置へ転送し、前記第一のサイトは、前記二重化、前記同期リモートコピー及び前記非同期リモートコピーの処理を再開することを特徴とする請求項 10 記載のデータ複製方法。

【請求項 13】

計算機及び記憶装置を有し、

前記計算機は、前記記憶装置の第一の記憶領域へデータベースのログを書き込み、前記記憶装置の第二の記憶領域へデータベースのデータを格納し、

前記記憶装置は、前記第一の記憶領域への更新データ及び前記第二の記憶領域への更新データを前記記憶装置と接続される第二の記憶装置へ同期リモートコピーで転送し、

前記計算機は、前記ログを該計算機と接続された第二の計算機へ転送することを特徴とする計算機システム。

【請求項 14】

第一、第二及び第三のサイトを有するシステムにおけるデータ複製方法であつ

て、

前記第一のサイトが有する計算機が、前記第一のサイトの記憶装置の第一の記憶領域へデータベースのログを書き込み、前記第一のサイトの記憶装置の第二の記憶領域へデータベースのデータを格納し、

前記記憶装置が、前記第一の記憶領域への更新データ及び前記第二の記憶領域への更新データを前記第二のサイトへ同期リモートコピーで転送し、

前記計算機が、前記ログを前記第三のサイトへ転送することを特徴とするデータ複製方法。

【請求項 15】

前記第一のサイトに障害が発生した場合、前記第二のサイトに格納された前記ログと前記第三のサイトに格納された前記ログの内容を一致させて、前記第二のサイトが有する計算機で前記第一のサイトの計算機で実行していた処理を継続することを特徴とする請求項 14 記載のデータ複製方法。

【請求項 16】

前記第一のサイトに障害が発生した場合、前記第二のサイトに格納された前記ログと前記第三のサイトに格納された前記ログの内容を一致させて、前記第三のサイトが有する計算機で前記第一のサイトの計算機で実行していた処理を継続することを特徴とする請求項 14 記載のデータ複製方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、記憶装置を含んだ情報処理システムに関し、さらに詳しくは複数の情報処理システム間で記憶装置に格納されたデータを転送する技術に関する。

【0002】

【従来の技術】

記憶装置を有する情報処理システムにおいて電源障害や天災等によって記憶装置に障害が発生した場合、当該情報処理システムを使用する業務が一時的に停止したり、最悪の場合、記憶装置に格納されたデータが失われることがある。このような事態を回避するために、当該情報処理システムとは異なる遠隔地に用意さ

れた記憶装置に、当該情報処理装置の記憶装置に格納されたデータを転送して複製する技術（以下「リモートコピー」と称する）が存在する。

【0003】

リモートコピーには同期リモートコピー及び非同期リモートコピーの二種類が存在し、各々長所と短所がある。具体的には、同期リモートコピーでは、情報処理システムの記憶装置は、情報処理システムの計算機からの書き込み要求があった場合、遠隔地に存在する記憶装置にその書き込み要求に付随するデータの転送が完了した後に、その書き込み要求に対する応答を計算機に行う。したがって、同期リモートコピーでは障害によるデータ消失が少ないが、記憶装置間の回線遅延が増加すると計算機と記憶装置間の I/O 性能が悪化する。

【0004】

一方、非同期リモートコピーでは、情報処理システムの記憶装置は、書き込み要求に対する計算機への返答と書き込み要求に付随するデータの遠隔地への転送とを独立したタイミングで実施する。したがって、非同期リモートコピーは記憶装置間の距離が長くても性能低下を招きにくい、データ消失の可能性は同期リモートコピーより高くなる。

【0005】

近年、双方のリモートコピーの短所を補うために、複数の情報処理システム（以下「サイト」）を用いたリモートコピーの技術が利用され始めている。

【0006】

例えば、特許文献 1 では、第一のサイトが有する記憶装置から近距離にある第二のサイト（第二の記憶装置を含む）及び第一のサイトから遠距離の第三のサイト（第三の記憶装置を含む）を含むシステムが開示されている。そして、本システムでは、二つのモードが交互に実行される。

【0007】

具体的には、第一のモードは、第一の記憶装置から第二の記憶装置へ同期リモートコピーが行われ、さらにリモートコピーされたデータは第二の記憶装置内で複製される。尚、このモードでは、第二の記憶装置から第三の記憶装置に対するリモートコピーは行われない。

【0008】

一方、第二のモードでは、第一の記憶装置から第二の記憶装置への同期リモートコピーを継続しつつ、第二の記憶装置から第三の記憶装置への非同期リモートコピーが行われる。ただし、このモードでは、第二の記憶装置内でのデータの複製は中止されている。

【0009】

また、非特許文献1でも、第一のサイトが有する記憶装置から近距離にある第二のサイト（第二の記憶装置を含む）及び第一のサイトから遠距離の第三のサイト（第三の記憶装置を含む）を含むシステムが開示されている。そして、本システムでは、第一の記憶装置から第三の記憶装置へのコピーを行うために二つのモードが交互に実行される。

【0010】

具体的には、常に第一の記憶装置から第二の記憶装置へ同期リモートコピーが行われている。又、第一のモードでは、第一のサイトの計算機が書き込んだデータが第一の記憶装置内で複製される。尚、このモードでは、第一の記憶装置から第三の記憶装置に対するリモートコピーは行われない。

【0011】

一方、第二のモードでは、第一の記憶装置から第三の記憶装置への非同期リモートコピーが行われる。ただし、このモードでは、第一の記憶装置内でのデータの複製は中止されている。

【0012】

【特許文献1】

米国特許第6209002号

【非特許文献1】

Claus Mikkelsenほか, "Addressing Federal Government Disaster Recovery Requirements with Hitachi Freedom Storage" (発行元 HITACHI DATA SYSTEMS、2002年11月発行)

【0013】**【発明が解決しようとする課題】**

特許文献1及び非特許文献1に開示された技術では、第三の記憶装置へのデータのコピーが定期的にしか行われないため、第一の記憶装置と第二の記憶装置が同時に障害になった場合に失われるデータが多くなる可能性がある。

本発明の目的は、第一の記憶装置と第二の記憶装置が同時に障害になった場合に失うデータ量を少なくする情報処理システムを提供することである。

【0014】**【課題を解決するための手段】**

本発明の一実施形態として、複数のサイトを有するシステムで、第一のサイトの複数の記憶領域にデータを二重化して格納し、そのうちの第一の記憶領域へのデータ更新の内容を第二のサイトへ同期リモートコピーで転送し、そのうちの第二の記憶領域へのデータ更新の内容を第三のサイトへ非同期リモートコピーを用いて転送する構成とする。

【0015】

更に、第一のサイトが障害等で使用不能となった場合に、第二のサイト又は第三のサイトで第一のサイトが行っていた処理を継続し、かつ他のサイトへ更新データを送信することで、データの複製を行う。

【0016】

更に、第一のサイトが復旧した場合、第二のサイト又は第三のサイトから第一のサイトへデータを転送し、その後、第一のサイトで同期及び非同期のリモートコピーを再開する構成も考えられる。

【0017】**【発明の実施の形態】**

図1は、本発明を適用した情報システムの第一の実施形態を示す図である。

情報システムは、3つのサイト、具体的には、Primaryサイト100、Syncサイト170及びAsyncサイト180を有する。各々のサイトは通信線160を介して相互に接続されている。尚、上述したように、サイトとは、計算機及び計算機と接続される記憶装置から構成される情報処理システムである。

。尚、サイト内の構成は以下の実施形態に限定されないのは言うまでもない。

【0018】

Primary サイト100は通常時にユーザが使用するサイトで、その計算機では、ユーザが使用するアプリケーションが実行される。

Sync サイト170は、Primary サイト100と地理的に異なる場所に存在するサイトである。Async サイト180は、Primary サイト100及びSync サイト170と地理的に異なる場所に存在するサイトである。尚、Primary サイト100とSync サイト170との間の距離は、Primary サイト100とAsync サイト180との間の距離よりも短いとして、以下説明する。

【0019】

Primary サイト100は、ホスト110、記憶装置120及びチャネルエクステンダ150を有する情報処理システムである。ホスト110は記憶装置120に対してデータの書き込み又は読み出し（以下「I/O」）を行う計算機である。記憶装置120は、ホスト110が用いるデータを保存する装置であり、ホスト110からのI/Oを処理する。尚、他のサイトも同様の構成を有する。

【0020】

ホスト110と記憶装置120は通信線130を通して接続されている。ここで通信線130としてはファイバーチャネルやATM、イーサネット（登録商標）等が考えられるが、ホスト110と記憶装置120の間でI/O処理を行うことができる媒体であればこれ以外であってもよい。

【0021】

又、記憶装置120とチャネルエクステンダ150は通信線140を介して接続されている。通信線140はファイバーチャネルで構成することが考えられるが、前述の目的を果たすことができればこれ以外であってもよい。

【0022】

チャネルエクステンダ150は、ネットワーク140を経由して記憶装置120から受信した情報を通信線160を経由して他サイトへ転送したり、通信線1

60を經由して他サイトから受信した情報を通信線140を通して記憶装置120へ転送するための装置である。尚、通信線140と通信線160が直接結合可能な場合は、チャンネルエクステンダ150は存在しなくてもよい。

【0023】

ホスト110は、OSやアプリケーション等のプログラムを実行するCPU111、メモリ112、通信線130を用いて記憶装置120とのI/Oを行うためのアダプタHBA(Host Bus Adapter)113を有し、これらは内部ネットワーク114で相互結合されている。ここで、CPU111やメモリ112、HBA113はホスト110内に複数存在してもよく、さらにホスト110はこれ以外の装置を有しても良い。また、内部ネットワーク114はI/Oバスやメモリバスで構成されることが考えられるが、本ネットワークに接続された機器を互いに通信させることが出来ればこれ以外の構成であってもよい。

【0024】

記憶装置120は、I/Oを処理したり、後述するアクセス制御やリモートコピーを実現するためのプログラムを実行するCPU121、メモリ122、CHA123、RCA124及び複数の記憶デバイス125を有する。

【0025】

CHA123は、記憶装置120が通信線130と接続するためのアダプタである。

RCA124(Remote Copy Adapter)は、記憶装置120から通信線140及びチャンネルエクステンダ150を介して他サイトの記憶装置と通信するためのアダプタである。

なお、CHA123とRCA124は一つのアダプタとして存在してもよく、また、通信線130、通信線140及び通信線160は互いが同一の通信線であってもよい。

【0026】

記憶デバイス125は、ホスト110から転送されたデータを保存するためのデバイスであり、例えばHDD、光磁気ディスク、CD-ROM、メモリディスク、Flash ROM等が考えられるがこれ以外ののものであってもよい。また、

高信頼化のために R A I D のような方法で複数の記憶デバイスをまとめて一つの仮想的な記憶デバイスとしてもよい。更に、前述の仮想的な記憶デバイスや元の記憶デバイスの領域を分割した後にこれを論理的な記憶デバイスとして取り扱っても良い。なお、ここには図示していないが、記憶装置 120 は記憶デバイス 125 に格納されるデータをキャッシングするためのキャッシュメモリを有してもよい。

また、図 1 では、通信線 160 はスター型接続の構成であるが、各サイト間で通信可能であればこれ以外のトポロジを採用してもよい。

【0027】

図 2 は、各サイトが有するプログラムやデータの構成を示す図である。これらのプログラムやデータは、それぞれの装置のメモリに格納される。尚、これらのプログラムによって各装置の C P U で実行される処理は、専用のハードウェアによって実現されてもよい。

【0028】

P r i m a r y サイト 100 のホスト 110 は、アプリケーションプログラム（以下「アプリケーション」）201、システム設定プログラム 211、ミラープログラム 212 及びミラー設定情報 213 を有する。

アプリケーション 201 は、具体的にはデータベースプログラムや W e b サーバプログラムである。ホスト 110 の使用者は、アプリケーション 201 を C P U 111 で実行させることで、記憶装置 120 にデータを書き込むことができる。尚、アプリケーション 201 は複数あってもよい。

【0029】

ミラープログラム 212 は、記憶装置 120 がホスト 110 に提供する一つ以上の記憶領域を用いて、ホスト 110 がデータの複製（以下「ミラーリング」）を記憶装置 120 に作成する際に C P U 110 で実行されるプログラムである。ミラー設定情報 213 には、ミラーリングを行うために必要な設定情報が含まれる。

ここで、記憶領域とは、記憶デバイス 125 に含まれる一つ以上のブロックまたはトラック等の単位から構成される領域であり、例としてはボリュームやパー

ション、スライスがある。

【0030】

Primary サイト100の記憶装置120は、アクセス制御プログラム221、同期リモートコピープログラム222、非同期リモートコピープログラム223及び記憶領域管理プログラム224を有する。

記憶領域管理プログラム224は、記憶装置120内の記憶デバイス125をホスト110がアクセス対象とする記憶領域として提供するための設定や管理を行う際にCPU121で実行されるプログラムである。ここで、記憶領域管理プログラム224は、記憶デバイス125の利用されていない領域の一部または全てを用いてホスト110に対して記憶領域を提供するための設定を記憶装置120が行ったり、既に設定された記憶領域を削除して再び利用されていないものとする設定を行ったり、既に設定された記憶領域に対してホスト110からどのような識別子でアクセスさせるかについての設定（パス設定）を行う際に実行される。

ここで、図2の記憶領域225Aと記憶領域225Bは、記憶領域管理プログラム224に基づいて作成された記憶領域である。

【0031】

なお、ホスト110が記憶領域にアクセスするために用いる識別子を外部記憶領域識別子と呼ぶ。この外部記憶領域識別子の例として、ファイバーチャネルを用いたシステムの場合はポート番号及びLUNの組が、ESCON（登録商標）やFICONを用いた場合はコントロールユニット番号及びデバイス番号の組が、IPネットワークを用いる場合はIPアドレスやポート番号が考えられるが、ホスト110が記憶領域にアクセスするための識別子として用いることが出来るものであればこれ以外の情報又は情報の組合せが識別子として用いられても良い。

【0032】

アクセス制御プログラム221は、記憶装置120内の記憶領域に対するホスト110からのアクセス要求を制御する際にCPU121で実行されるプログラムである。

【 0 0 3 3 】

同期リモートコピープログラム 2 2 2 は、記憶装置 1 2 0 内の記憶領域と記憶装置 1 2 0 とは別な記憶装置内に存在する記憶領域との間で同期リモートコピーを行う際に、CPU 1 2 1 で実行されるプログラムである。以後の説明では、コピー元となる記憶領域を正記憶領域、コピー先となる記憶領域を副記憶領域と呼ぶことにする。また、正記憶領域を有する記憶装置を正記憶装置、副記憶領域を有する記憶装置を副記憶装置と呼ぶことにする。

【 0 0 3 4 】

同期リモートコピーが実行される際、二つの記憶領域のデータおよびコピーの状況を示したり操作するために、リモートコピーのペアである正副記憶装置は、複数のペア状態（Simplex、Initial-Copying、Duplex、Suspend 及び Duplex-Pending）を示す情報を管理する。尚、ペア状態の情報には、お互いの記憶領域がリモートコピーのペア関係であることを示す情報も含まれる。

【 0 0 3 5 】

Simplex 状態は、正副の記憶領域間での同期リモートコピーが開始されていない状態である。Duplex 状態は、同期リモートコピーが開始され、後述する初期化コピーも完了して正副の記憶領域の内容が同一となった状態である。同期リモートコピーの場合、正記憶領域に対して行われた書き込みの内容が副記憶領域に対しても反映された後に、書き込みを行ったホスト 1 1 0 に対して正常完了のメッセージが返される。従って、書き込み途中の記憶領域を除けば、正記憶領域のデータと副記憶領域のデータの内容は同じとなる。

【 0 0 3 6 】

ただし、例えば記憶領域毎に一意的な識別子が保存される場合等では、記憶領域の特定の部分について正副の記憶領域の内容が同一でない場合があっても良いものとする。なお、こうしたデータの同一性を、以降の説明では巨視的に同一であるという表現する。また、ペアが Duplex 状態中に、例えば正副記憶領域でのデータ同一性を保持するため等の目的で副記憶領域に対する書き込み要求を拒否あるいはこれに類する処理を行っても良いものとする。

【0037】

Initial-Copying 状態は、Simplex 状態から Duplex 状態へ遷移するまでの中間状態であり、この期間中に、必要ならば正記憶領域から副記憶領域への初期化コピー（正記憶領域に既に格納されていたデータのコピー）が行われる。初期化コピーが完了し、Duplex 状態へ遷移するために必要な処理が終わった段階でペア状態は Duplex となる。

【0038】

Suspend 状態は、正記憶領域に対する書き込みの内容を副記憶領域に反映させない状態である。この状態では、正副の記憶領域のデータの巨視的な同一性は保障されなくなる。オペレータやホスト、又は記憶装置を管理する計算機（図示せず）の指示を契機に、ペア状態は、他の状態から Suspend 状態へ遷移する。それ以外に、正記憶領域から副記憶領域への同期リモートコピーを行うことが出来なくなった場合に自動的にペア状態が Suspend 状態に遷移することが考えられる。

【0039】

以後の説明では、後者の場合を障害 Suspend 状態と呼ぶことにする。障害 Suspend 状態となる代表的な原因としては、正副の記憶領域の障害、正副の記憶装置の障害、正副間の通信路障害が考えられる。なお、Duplex 状態の副記憶装置で副記憶領域に対する書き込み要求を拒否またはそれに類する処理を行っていた場合は、副記憶装置は、障害 Suspend 状態で書き込み処理を許可してもよい。さらに正副記憶装置は、Suspend 状態となった以降の正副の記憶領域に対する書き込み位置を記録してもよい。

【0040】

また、正記憶装置は、障害 Suspend 状態時に正記憶領域に対する書き込みを拒否しても良い。こうすると、正記憶装置と副記憶装置との間の通信路が切断された後でも正副の記憶装置のデータが同一であるため、切断後に正記憶装置に障害が発生した場合でもデータ消失を防ぐことが出来る。

【0041】

Duplex-Pending 状態は、Suspend 状態から Duplex

状態に遷移するまでの中間状態である。この状態では、正記憶領域と副記憶領域のデータの内容を巨視的に一致させるために、正記憶領域から副記憶領域へのデータのコピーが実行される。正副記憶領域のデータの同一性を確保できた後にペア状態はDuplexとなる。なお、Duplex-Pending状態におけるデータのコピーは、Suspend状態で正副記憶装置が記録した書き込み位置を利用して更新が必要な部分だけをコピーする差分コピーが用いられても良い。

【0042】

なお、以上の説明ではInitial-Copying状態とDuplex-Pending状態は別々な状態としたが、これらをまとめて一つの状態として管理装置の画面に表示したり、状態を遷移させても良い。

【0043】

非同期リモートコピープログラム223は、記憶装置120内の記憶領域と記憶装置120とは別な記憶装置内に存在する記憶領域との間で非同期リモートコピーを実行する際に、CPU121で実行されるプログラムである。既に説明したように、非同期リモートコピーでは、Duplex状態の副記憶領域に対する書き込みデータの反映は、記憶装置120のホスト110に対する書き込みの正常完了メッセージの送信とは無関係（非同期）に行われる。

【0044】

非同期リモートコピーの場合、正記憶領域から副記憶領域へのデータコピーの方法として、以下の方法がある。

例えば、正記憶装置が、データが書き込まれた記憶領域のアドレスを含んだ制御情報及び書き込まれたデータの組（以下「ログエントリ」）を、データの書き込みの度に作成し、これを副記憶装置へ転送し、副記憶領域へ反映させる方法がある。さらにこの発展形として、ログエントリの制御情報に書き込みの時間順序を示す情報を含め、副記憶領域へログエントリを反映させる際にはこの時間順序を示す情報を利用して時間順序どおりに反映する方法がある。

【0045】

また、本方法の効率的な方法として、正記憶領域の同一領域に対する書き込み

が連続して発生した場合には、正記憶装置は、途中の書き込みに対するログエントリは副記憶装置へ転送せず、最後の書き込みに対するログエントリのみを転送する方法がある。また、正記憶装置がキャッシュメモリを有する場合、正記憶装置がキャッシュメモリ上に書き込まれたデータを示すポインタをログエントリの制御情報に付加し、キャッシュメモリ上のデータが別な書き込み処理によって更新されるまではログエントリ作成用のデータコピーを遅延させる方法もある。

【 0 0 4 6 】

非同期リモートコピーが実行される際にも、正副記憶装置は、ペア状態（Simplex、Initial-Copying、Duplex、Suspend、Duplex-Pending及びSuspending）を管理する。Simplex、Initial-Copying、Suspend及びDuplex-Pending状態については同期リモートコピーと同様である。

【 0 0 4 7 】

Duplex状態も基本的には同期リモートコピーの場合と同じであるが、書き込みデータの副記憶領域への反映が非同期に行われるため、データの同一性は同期リモートコピーとは異なる。

Suspending状態とは、Duplex状態からSuspend状態へ遷移するまでの中間状態であり、非同期リモートコピーの場合は、Suspending状態を経由してSuspend状態へ遷移する。尚、この状態で、正副記憶装置は、両記憶装置のメモリで保持されているログエントリを副記憶装置へ反映させる処理を行っても良い。また、同期リモートコピーのSuspend状態で述べた正副の記憶領域に対する書き込み位置の記録を行っている場合は、正副記憶装置は、反映できなかったログエントリを書き込み位置の記録に加える。

【 0 0 4 8 】

SynCサイト170の記憶装置272は、同期リモートコピープログラム22及び非同期リモートコピープログラム223を有する。また、図示はしていないが、記憶装置272も記憶領域管理プログラム224を有し、記憶領域225Cはこのプログラムに基づいて作成された記憶領域である。なお、記憶装置272にはこれ以外のプログラム（例えばアクセス制御プログラム221）が含ま

れていても良い。

【0049】

A s y n c サイト 180 の記憶装置 282 は、非同期リモートコピープログラム 223 を有する。記憶装置 272 の場合と同様に、記憶装置 282 も記憶領域管理プログラム 224（図示せず）を有し、記憶領域 225 D は、このプログラムに基づいて作成された記憶領域である。また、記憶装置 282 にはこれ以外のプログラム（例えばアクセス制御プログラム 221）が含まれていても良い。

【0050】

ここで、記憶領域 225 A、記憶領域 225 B、記憶領域 225 C 及び記憶領域 225 D は、リモートコピーで使用する際には、すべて同じ記憶容量であることが望ましい。ただし、リモートコピーを行うことができるのであれば、各記憶領域の容量が異なってもよい。

【0051】

以下、簡単に本実施形態での動作を説明する。尚、以下プログラムが主語になる場合は、各プログラムが格納されている装置の C P U が実際の処理を行っているものとする。更に、プログラム間でデータが遣り取りされる記載は、実際には、計算機で使用するプログラム間通信（共通のメモリ等を介してプログラムが同一のデータを扱う）が実行されているものとする。

【0052】

本実施形態では、ホスト 110 上のアプリケーション 201 で生成された書き込み用のデータを、ホスト 110 上のミラープログラム 212 が複製して、記憶装置 120 が有する二つの記憶領域に送信する（ミラーリング）。

【0053】

P r i m a r y サイトの正記憶装置においては、同期リモートコピープログラム 222 が、二つの記憶領域のうち一方の記憶領域を用いてコピー処理を行い、記憶領域に格納されたデータを S y n c サイトの記憶領域へ同期リモートコピーする。又非同期リモートコピープログラム 223 が、もう一方の記憶領域を用いてコピー処理を行い、もう一方の記憶領域に格納されたデータを A s y n c サイトの記憶領域に非同期リモートコピーする。

【0054】

更に、Primaryサイトが被災した場合は、SyncサイトまたはAsyncサイトのどちらかでアプリケーションの処理を再開する。Syncサイトの場合、同期リモートコピーによってコピーされた記憶領域が用いられる。一方、Asyncサイトの場合は、非同期リモートコピーによってコピーされた記憶領域が用いられる場合と、さらにSyncサイトから記憶領域の差分データをAsyncサイトにコピーしてからアプリケーションを再開する場合がある。

【0055】

以下、本実施形態における処理の各詳細について説明する。

まず、ホスト110のCPU111が実行するミラーリングについて説明する。CPU111は、ミラープログラム212を実行することで、記憶装置120が有する二つ以上の記憶領域に同一のデータを格納し、その複数の記憶領域を1つの仮想的な記憶領域（以下「仮想記憶領域」）としてアプリケーションに提供する。

【0056】

尚、PrimaryサイトとSyncサイト間で通信路障害が発生した場合、記憶装置120はホスト110に対して記憶領域225Aに対する書き込みを拒否することができる。この場合、この拒否通知を受け取ったミラーリングプログラム212は、記憶領域225Bへの書き込みも行わずにアプリケーション201に書き込み失敗を通知しても良い。なお、以後の説明ではミラーリングに使用される一つ以上の記憶領域の集合をミラーグループと呼ぶことにする。

【0057】

図8は、ホスト110が仮想記憶領域（ミラーグループから構成される）を管理するためのミラー設定情報213の内容を示すブロック図である。

ミラー設定情報213は、1つの仮想記憶領域の設定に関する情報を登録するエントリ810を仮想記憶領域の数だけ有する。各々のエントリ810は、仮想記憶領域のミラーグループに所属する記憶領域に割り当てられている外部記憶領域識別子を登録するフィールド801、同期状態情報を登録するフィールド802、障害状態を登録するフィールド803及び保護設定情報を登録するフィール

ド 804 を有する。

【0058】

同期状態情報は、ミラーグループに新たに記憶領域を追加した場合のデータの同期状態を示す情報である。ミラーグループに追加された直後の記憶領域には未だミラーグループに属する記憶領域のデータが複製されていないので、この記憶領域に対する同期状態情報は同期が取れていないことを示す情報となる。その後、既にミラーグループに所属している記憶領域からこの新しい記憶領域へデータのコピーが完了して他の記憶領域のデータと同期が取れた状態になった段階で、この記憶領域に対応する同期状態情報は同期が取れたことを示す情報となる。また、障害状態となった記憶領域に対しても同期が取れていないことを示す情報が設定される。

【0059】

障害状態情報は、エントリに対応する記憶領域の障害状態を示す情報である。ここで、障害状態には、記憶領域が利用不可能になった状態のほか、同期リモートコピーの障害 *Suspend* 書き込み禁止の設定によって記憶領域が書き込み禁止になった状態も含まれる。

【0060】

保護設定情報は、ミラーグループに所属するいずれかの記憶領域が同期リモートコピーの障害 *Suspend* 書き込み禁止の設定によって書き込み禁止となった場合に、その記憶領域へのアプリケーション 201 からの更新を拒否することを示す情報である。

【0061】

図 9 は、ミラーグループに所属する記憶領域への書き込み処理のフローを示す図である。本処理は、正記憶装置から他サイトへのデータの転送の如何に関わらず、システムの利用者等の指示によって開始される。

ホスト 110 は、アプリケーション 201 で発生した仮想記憶領域へのデータの書き込み要求の処理を開始する（ステップ 901）。

書き込み要求の処理を開始したら、ホスト 110 は、ミラーグループに所属している各記憶領域に対応するフィールド 803 の障害状態情報をチェックし、障

害 S u s p e n d 書き込み禁止の設定によって書き込み禁止になっている記憶領域が存在するかどうかチェックする（ステップ 906、907、908、909）

書き込み禁止になっている記憶領域が一つでも存在した場合、ホスト 110 は、アプリケーション 201 に対して書き込みの異常終了を報告し、処理を終了する。なお、報告には障害 S u s p e n d 書き込み禁止が異常終了の原因であることをあわせて報告してもよい（ステップ 910）。

【0062】

書き込み禁止になっている記憶領域が存在しない場合、ホスト 110 は、ミラーグループに所属する記憶領域を指し示すために用いる変数 i と書き込みが成功した記憶領域の数を示すために用いる変数 j の値をゼロに初期化する（ステップ 902、911）。

【0063】

その後、ホスト 110 は、ミラーグループに所属している各記憶領域に対して、まずフィールド 803 に登録されている障害状態情報をチェックして、障害状態でなければデータの書き込みを行う。また、障害状態であれば変数 i をインクリメントして次の記憶領域に対する処理へ移る（ステップ 912、913、903、915）。

【0064】

ホスト 110 は、ステップ 915 でのデータの書き込み結果をチェックし、書き込み成功ならば変数 i と j をインクリメントして次の記憶領域に対する処理へ移る。失敗ならば、当該記憶領域に対応するフィールド 803 の障害状態情報を書き換えて、障害が発生したことを記録する（ステップ 916、914、917）。

【0065】

その後、ホスト 110 は、書き込み失敗の理由が障害 S u s p e n d 書き込み禁止によるものならば、障害 S u s p e n d 書き込み禁止となっている記憶領域から既に書き込みが完了した記憶領域へ、書き込みの対象となったアドレス領域に格納されたデータのコピーを記憶装置 120 に指示し、ミラーグループに属す

る記憶領域のデータをデータ書き込み前の状態に戻す。なお、障害 S u s p e n d 書き込み禁止に遭遇する前に既に更新してしまった記憶領域にコピー先を限定してコピーの効率化を図っても良い（ステップ 9 1 8、9 2 0）。

【 0 0 6 6 】

ステップ 9 2 0 の処理後、ミラープログラム 2 1 2 は、アプリケーション 2 0 1 に対して書き込みの異常終了を報告し、処理を終了する。なお、報告には障害 S u s p e n d 書き込み禁止が原因であることをあわせて報告してもよい（ステップ 9 2 1）。

【 0 0 6 7 】

ステップ 9 1 2 でミラーグループに所属する全ての記憶領域に対する処理が終わったと判断したら、ホスト 1 1 0 は、変数 j をチェックして書き込みが正常終了した記憶領域の数が一つ以上存在するか確認する。もし存在すればミラープログラム 2 1 2 は、アプリケーション 2 0 1 に対して正常終了を報告し、そうでなければ異常終了を報告する（ステップ 9 1 9、9 2 2、9 2 3）。

【 0 0 6 8 】

以上の処理を行って、ホスト 1 1 0 は、ミラーグループへのデータの更新をする。又、以上の処理を行うことで、ミラープログラム 2 1 2 は、副記憶装置の記憶領域 2 2 5 C に対して更新データが反映されない記憶領域がミラーグループに存在する場合には、ミラー処理中にアプリケーションに対して書き込み異常終了を返す、即ちミラーグループの他の記憶領域へのデータ更新を行わないことができる。この様な処理を障害保護処理といい、障害保護処理を行う設定を障害保護設定と称する。尚、上述のステップで、障害保護設定を行わない、即ちミラーグループのある記憶領域へのデータの書き込みが許可されなくても、他の記憶領域へのデータの書き込みを許可する構成でも良い。

【 0 0 6 9 】

したがって、P r i m a r y サイトから S y n c サイトへの通信路障害等が原因である障害 S u s p e n d が発生した後に P r i m a r y サイトが障害停止したとしても、記憶領域 2 2 5 C と記憶領域 2 2 5 D のデータ内容が同一となる（記憶領域 2 2 5 A と 2 2 5 B の内容の一致が保証されるから）ため、以後に述べ

る復旧処理で記憶領域 2 2 5 C と記憶領域 2 2 5 D 間でのリモートコピーのペアを設定を行う際の初期化コピーを省略することができる。

【0 0 7 0】

図 6 は、ホスト 1 1 0 における、ミラーグループに新たに記憶領域を追加する場合の処理手順を示す図である。なお、記憶領域の追加指示（外部記憶領域識別子の情報を含む）はホスト 1 1 0 の使用者、ホスト 1 1 0 上のプログラム、管理計算機からミラープログラム 2 1 2 へ行われるが、これ以外の計算機が追加指示をミラープログラム 2 1 2 へ出してもよい。また、追加指示はホスト 1 1 0 の使用が開始される時点で行われることが考えられるが、本発明はこれに限定されない。

【0 0 7 1】

まずホスト 1 1 0 は、追加対象である記憶領域に割り振られた外部記憶領域識別子を追加先のミラーグループに対応するエントリ 8 1 0 のフィールド 8 0 1 に追加する。また、追加したフィールド 8 0 1 に対応するフィールド 8 0 2 には同期がとれていないことを示す情報を、フィールド 8 0 3 には正常状態を示す情報を書き込む（ステップ 6 0 1）。

【0 0 7 2】

その後、アプリケーション 2 0 1 で書き込み要求が発生した場合、ホスト 1 1 0 は、ミラーグループ内の他の記憶領域と追加対象である記憶領域の両方にデータを書き込むようにする。

又、追加された記憶領域と既存のミラーグループに格納されていたデータとが同期が取れるまでは、ホスト 1 1 0 は、アプリケーション 2 0 1 で読み込み要求が発生した場合は、追加対象である記憶領域ではなく、ミラーグループ内の既に同期状態である記憶領域からデータを読み出す（ステップ 6 0 2）。

【0 0 7 3】

その後、ホスト 1 1 0 は、ミラーグループ内の既に同期状態である記憶領域から追加対象である記憶領域へのデータコピーを記憶装置 1 2 0 へ指示する。なお、データコピーは記憶領域全体を対象とすることが考えられるが、事前にコピーの必要が無い領域がビットマップ等でわかる場合は、当該領域のコピーを行わな

くても良い（ステップ604）。

追加対象の記憶領域へのデータのコピーの完了後、ホスト110は、追加対象の記憶領域のフィールド802の情報を同期状態を示す情報に更新する（ステップ605）。

【0074】

その後、アプリケーション201で読み込み要求が発生した場合、ホスト110は、追加対象である記憶領域も含めたミラーグループ内の既に同期状態である記憶領域からデータを転送する（ステップ606）。

【0075】

次に、本実施形態におけるシステムの同期、非同期リモートコピーの初期設定の処理（以下「初期化処理」）について説明する。なお、本処理はホスト110のプログラムに基づいてCPU111で実行されるが、ホスト110の管理者が直接記憶装置へ設定情報を入力することで実行されたり、それ以外の計算機によって実行されてもよい。

図3は、初期化処理の処理手順を示す図である。なお、システムの初期状態として、記憶領域225Aは既に作成済み及びパス設定済みで、ミラー設定情報213には記憶領域225Aが単独でミラーグループを組むことを示す情報が登録されている状態を想定する。以下、正記憶装置は、図3の左側の処理ステップと右側の処理ステップを並行して行う。

【0076】

ホスト110は、記憶領域225Bの作成を正記憶装置に指示し、パス設定やアクセス制限設定等の設定を行う。好ましい形態としては、記憶領域225Bのアクセス制限設定は記憶領域225Aと同じにする。しかし、記憶領域225Bがホスト110からアクセス可能であればこれ以外のアクセス制限設定を行っても良い。

その後、ホスト110は、先述した処理手順に従ってミラーグループに記憶領域225Bを追加する（ステップ304）。

【0077】

その後、正記憶装置は、ホスト110の指示に従って非同期リモートコピー

プログラム 223 の実行を開始し、記憶領域 225 B から A s y n c サイトに存在する記憶領域 225 D に対して非同期リモートコピーにおける初期化コピーを開始し、そのペア状態が D u p l e x になるまで待つ（ステップ 305）。

【0078】

一方、正記憶装置は、ホスト 110 の指示に従って同期リモートコピープログラム 222 を実行して、記憶領域 225 A から S y n c サイトに存在する記憶領域 225 C に対して同期リモートコピーにおける初期化コピーを開始し、ペア状態が D u p l e x になるまで待つ（ステップ 310）。

その後、正記憶装置は、ステップ 305 とステップ 310 の両方の処理が終わるのを待つ（ステップ 306）。

【0079】

尚、上記の処理では、既に作成されている記憶領域 225 A を同期リモートコピーの正記憶領域として用いたが、新たに追加した記憶領域 225 B が同期リモートコピーの正記憶領域として用いられても良い。この場合、記憶領域 225 A と記憶領域 225 B に対する操作を入れ替えるだけで、後は上述の手順と同様でよい。さらに、本処理の前に予め記憶領域 225 B を作成してミラーグループに登録しておき、ステップ 302 から 304 の処理を省略しても良い。

【0080】

又、初期化処理におけるホスト 110 からの同期又は非同期リモートコピーの開始指示には、正記憶領域、副記憶領域、正記憶装置、副記憶装置を指定する情報、すなわちリモートコピーのペアを指定する情報が含まれている。あるいは別の実施形態として、ホスト 110 又は管理用の計算機からペアを指定する情報が予め正副記憶装置へ送信され、リモートコピーの開始指示には、そのペアを指定する識別子だけが含まれていても良い。

以上の処理により、本システムにおける同期、非同期リモートコピーの実施（以下「通常運用」）準備が整う。

【0081】

次に、本実施形態のシステムの通常運用時における動作を説明する。

図 4 は、システムの動作におけるデータの移動を概念的に示す模式図である。

なお、本図において、矢印 4 0 1 から矢印 4 0 5 はホスト 1 1 0 から送信された書き込み要求に対応するデータの流れを示すものである。尚、本図では、既に図 3 で示した初期化処理が済んでいるとする。

【0 0 8 2】

まず、ホスト 1 1 0 のアプリケーション 2 0 1 で仮想記憶領域のデータの更新が要求されると、ホスト 1 1 0 はミラープログラム 2 1 2 の実行を開始して、更新用のデータ（以下「更新データ」）の処理を開始する（矢印 4 0 1）。

ホスト 1 1 0 は、仮想記憶領域に含まれるミラーグループ全てに更新データを送信するため、更新データを複製して、矢印 4 0 2 と矢印 4 0 3 のように記憶領域 2 2 5 A と記憶領域 2 2 5 B へ転送する。転送順序は問わない。又、実際のデータの更新の手順は、上述した通りである。

【0 0 8 3】

正記憶装置は、記憶領域 2 2 5 A に対する更新データを受信した場合、更新データを記憶領域 2 2 5 A に反映した上で、その更新データを記憶領域 2 2 5 C に対して転送する。尚、記憶領域 2 2 5 A への反映は、実際には正記憶装置が有するキャッシュメモリへの書き込みの完了であっても良い（矢印 4 0 5）。

記憶領域 2 2 5 C に対する更新データの反映が終わったら、正記憶装置は、ホスト 1 1 0 に対して記憶領域 2 2 5 A に対する書き込みの正常終了メッセージを返す。

【0 0 8 4】

一方、正記憶装置は、記憶領域 2 2 5 B に対して更新データが転送された場合、記憶領域 2 2 5 B へ更新データを書き込み、ホスト 1 1 0 に対して記憶領域 2 2 5 B に対する書き込みの正常終了メッセージを返す。その後、正記憶装置は、記憶領域 2 2 5 B に対して書き込まれた更新データを、正常終了メッセージと非同期に記憶領域 2 2 5 D に対して転送する。（矢印 4 0 4）

記憶領域 2 2 5 A と記憶領域 2 2 5 B の両方に対する更新データの書き込みの書き込み正常終了メッセージを受信したミラープログラム 2 1 2 は、アプリケーション 2 0 1 に対して書き込み正常終了メッセージを返す。

なお、ホスト 1 1 0 から記憶領域 2 2 5 A 及び記憶領域 2 2 5 B への更新デー

タの転送は、互いに独立なタイミングで行われても良い。

【0085】

次に、本実施形態において、Primaryサイト100で障害等によるシステム停止が発生した場合の他サイトでのアプリケーションの再開手順、及びPrimaryサイト100が復活した場合の復帰手順について説明する。

尚、ホスト110におけるミラーリングが保護情報設定がオンで行われていれば、Primaryサイト100の記憶装置120とSyncサイトの記憶装置272の間の通信が停止した後にPrimaryサイト100が障害で停止した場合でも、ホスト110は、アプリケーション201に対して書き込み完了を通知した更新データを消失することはない。これは記憶装置間の通信が出来なくなった状況では、ホスト110は、アプリケーションからの書き込み処理を停止させるからである。

【0086】

図10は、Primaryサイト100がシステム停止になった際に、Primaryサイト100で実行されていたアプリケーションを、Syncサイト170で再開する手順を示す図である。

記憶領域225Aから記憶領域225Cに対する同期リモートコピーのペア状態によって記憶領域225Cに対して書き込みができない状態ならば、Syncサイト170のホスト271は、記憶装置272が有する同期リモートコピーのペア状態を変更し、記憶領域225Cへの書き込みを可能にする。なお、前述のペア状態の情報の変更としてはペア状態をSimplexに遷移させることが考えられる（ステップ1001、1002）。

【0087】

その後、ホスト271にてアプリケーション201を起動し、記憶領域225Cを用いて処理を再開する（ステップ1003）。

次に、ホスト281が、記憶領域225Bから記憶領域225Dに対する非同期リモートコピーのペア状態をSimplexに遷移させるように、記憶装置282に指示する（ステップ1004）。

【0088】

この際、記憶装置 2 7 2 と記憶装置 2 8 2 との間で通信可能であれば、ホスト 2 7 1 は、記憶領域 2 2 5 C から記憶領域 2 2 8 D への非同期リモートコピープログラム 2 2 3 によるペア設定及び非同期リモートコピーの開始を記憶装置 2 7 2 へ指示し、ペア状態が Duplex になるまで待つ（ステップ 1 0 0 5、1 0 0 6、1 0 0 7）。

【0 0 8 9】

これにより、Sync サイトでアプリケーションが再起動し、かつ、そのアプリケーションによる更新データが Async サイトへ転送される。

【0 0 9 0】

図 5 は、図 1 0 の処理手順の結果、Sync サイト 1 7 0 でアプリケーションが再開された後のシステム状態を示したブロック図である。

Sync サイト 1 7 0 のホスト 2 7 1 が記憶領域 2 2 5 C に対して発行したデータの更新（矢印 5 0 1）は、記憶装置 2 7 2 が非同期リモートコピーを実行することで、非同期に記憶領域 D 2 2 8 に対して反映される（矢印 5 0 2）。

【0 0 9 1】

したがって、Sync サイト 1 7 0 でアプリケーションの実行を再開した後で、Sync サイト 1 7 0 が災害停止した場合でも、Async サイト 1 8 0 の記憶領域 2 2 5 D を用いることで、Sync サイトで更新された最近のデータについては消失の可能性があるものの、それ以外のデータについては消失を回避することが可能となる。

図 1 1 は、Sync サイト 1 7 0 でアプリケーション処理を再開した状態から、通常運用状態（Primary サイト 1 0 0 の運用）への復帰手順を示す図である。

まず、Primary サイト 1 0 0 を復旧させる。この際、ホスト 1 1 0 には、記憶領域 2 2 5 A のミラー設定情報 2 1 3 の設定が施されておらず、記憶装置 1 2 0 においてペア情報も存在しないものとする（ステップ 1 1 0 1）。

その後、ホスト 2 7 1 かホスト 2 8 2 が、記憶装置 2 7 2 か記憶装置 2 8 2 に対して指示することで、記憶領域 2 2 5 C から記憶領域 2 2 5 D に対する非同期リモートコピーのペア状態を Simplex に遷移させる（ステップ 1 1 0 2）

。

【0 0 9 2】

その後、記憶装置 2 7 2 は、ホスト 2 7 1 の指示に基づいて記憶領域 2 2 5 C から記憶領域 2 2 5 A に対して同期リモートコピーを開始し、ペア状態が Duplex になるまで待つ（ステップ 1 1 0 3、1 1 0 4）。

次に、使用者又は管理者が、ホスト 2 7 1 上のアプリケーション 2 0 1 を停止する（ステップ 1 1 0 5）。

【0 0 9 3】

その後、ホスト 2 7 1 またはホスト 1 1 0 は、ステップ 1 1 0 3、1 1 0 4 で作成されたペアの正副関係を入れ替え、記憶領域 2 2 5 A から記憶領域 2 2 5 C に対して同期リモートコピーを行うよう、記憶装置 1 2 0 に指示する。もし、同期リモートコピープログラム 2 2 2 でこのような処理が実行できない場合、使用者は、一旦記憶領域 2 2 5 A と 2 2 5 C のペア状態の情報を削除してから、逆方向にペアを設定することになる。その場合、既に記憶領域 2 2 5 A と記憶領域 2 2 5 C は同じデータを有するため、逆方向の同期リモートコピーの初期化コピーは省略されてもよい（ステップ 1 1 0 6）。

【0 0 9 4】

この時点では、ホスト 1 1 0 には記憶領域 2 2 5 A を用いるための設定情報が登録されていないため、使用者は、記憶領域 2 2 5 A を含めたミラーグループを仮想記憶領域としてアプリケーション 2 0 1 に提供する設定を行う（ステップ 1 1 0 7）。

その後、使用者は、ホスト 1 0 0 上のアプリケーション 2 0 1 を再起動し、記憶領域 2 2 5 A を用いた処理を再開する（ステップ 1 1 0 8）。

【0 0 9 5】

ホスト 1 1 0 は、ステップ 1 1 0 7 で設定されたミラーグループに記憶領域 2 2 5 B を追加する（ステップ 1 1 0 9）。

記憶装置 1 2 0 は、ホスト 1 1 0 の指示に従って、記憶領域 2 2 5 B から記憶領域 2 2 5 D に対する非同期リモートコピーを開始する（ステップ 1 1 1 0）。

その後、記憶装置 1 2 0 は、記憶領域 B 2 2 6 のデータが記憶領域 2 2 5 A の

データ内容と一致し、ステップ1110のペア状態がDuplexになるまで待つ。

【0096】

尚、上述した手順において、ステップ1105のアプリケーション201の停止は、このステップより前に実行されてもよく、ステップ1108のアプリケーション201の再起動は、このステップよりも後に実行されてもよい。

上述の処理によって、Primaryサイトにおける処理を再開し、かつ、同期及び非同期リモートコピー間のデータの整合性も復旧することが出来る。

【0097】

図12は、Syncサイト170でアプリケーション処理を再開した状態から、通常運用状態（Primaryサイト100の運用）への復帰手順の別例を示す図である。尚、以下の処理において、Primaryサイトには既に記憶領域225A及びBが設定されている（ただしミラーグループの設定は未だ）とする。

まず、使用者は、ホスト271でのアプリケーション201の実行を停止する（ステップ1201）。

【0098】

ホスト281及びホスト271は、記憶領域225C及び記憶領域225Dが有するデータの内容が同期したことを確認して、ペア状態の情報の削除を記憶装置282及び272へ指示する（ステップ1202、1203）。

ホスト271は、記憶領域225Cから記憶領域225Bにデータをコピーするよう、記憶装置272へ指示する（ステップ1204）。

その後、記憶装置120は、ホスト110等の指示に基づいて記憶領域225Bから記憶領域225Dに対する非同期リモートコピーを開始する。なお、記憶領域225Bと記憶領域225Dには既に同じデータが格納されているため、非同期リモートコピーにおける初期化コピーは省略されてもよい（ステップ1205）。

【0099】

この時点で、ホスト110には、記憶領域225Bを用いるためのミラー設定

情報が存在しないため、使用者は、記憶領域 225 B を含めたミラーグループを仮想記憶領域としてアプリケーション 201 に提供する設定を行う（ステップ 1206）。

使用者は、ホスト 110 のアプリケーション 201 を再起動し、記憶領域 225 B を用いた処理を再開させる（ステップ 1207）。

その後、使用者は、ステップ 1206 のミラーグループに記憶領域 225 A を追加する（ステップ 1208）。

【0100】

その後、記憶装置 120 は、ホスト 110 の指示に従い、記憶領域 225 A から記憶領域 225 C に対する同期リモートコピーを開始する（ステップ 1209）。

ホスト 110 は、記憶領域 225 A のデータが記憶領域 225 B のデータ内容と一致し、ステップ 1209 のペア状態が Duplex になるまで待つ（ステップ 1210、1211）。

【0101】

尚、ステップ 1207 のアプリケーションの再起動をステップ 1209 の後にすることで、ステップ 1209 の初期化コピーを省略することができる。その場合はさらに、ステップ 1204 でホスト 110 が記憶領域 225 A となる記憶領域へのコピーを記憶領域 225 B と共に記憶装置 272 へ指示し、ステップ 1208 の記憶領域 225 A 追加時のミラーグループ作成時のコピーを省略してもよい。

上述した二つの処理によって、Primary サイトが復旧した際に、システムでは非同期リモートコピーからでも、同期リモートコピーからでも先に再開することができる。

【0102】

図 13 は、Primary サイト障害時に、Async サイト 180 でアプリケーションの実行を再開する手順を示す図である。

記憶装置 282 が有する記憶領域 225 B から記憶領域 225 D に対する非同期リモートコピーのペア状態情報が記憶領域 225 D に対して書き込みが可能な

いことを示す状態ならば、ホスト 281 は、記憶装置 282 へペア状態の情報の変更を指示し、記憶領域 225D への書き込みを可能にする。なお、ペア状態の情報の変更の方法の例としてはペア状態を Simplex に遷移させることが考えられる（ステップ 1301、1302）。

【0103】

ホスト 281 にてアプリケーション 201 が起動され、記憶領域 225D を用いて処理が再開される。なお、再開に際して、アプリケーション 201 による復旧処理を行っても良い。具体的には、データベースの場合ならば（1）DB にライトバック方式のバッファが存在することによる不整合が発生するため、トランザクションログを用いて修復する、（2）コミットされていないトランザクションの書き込みをトランザクションログを使ってトランザクション前の状態に戻す処理等である（ステップ 1303）。

これにより、Async サイトでアプリケーションの実行が再開できる。

【0104】

図 14 は、Primary サイト 100 が障害等で停止した際に、Sync サイト 170 の記憶領域 225C のデータを利用して Async サイト 180 上でアプリケーション 201 を再開する手順例を示す図である。

ホスト 271 及びホスト 281 は、記憶装置 272 と記憶装置 282 との間が通信可能か確認し、通信不可能であれば、本手順ではなく、例えば前述の Async サイト 170 での再開手順を行う（ステップ 1401）。

【0105】

ホスト 271 及びホスト 281 は、記憶領域 225A から記憶領域 225C に対する同期リモートコピーのペア状態の情報と、記憶領域 225B から記憶領域 225D に対する非同期リモートコピーのペア状態を Simplex に遷移させるように記憶装置 272 及び 282 へ指示する（ステップ 1402、1403）。

その後、ホスト 271 又はホスト 281 は、記憶領域 225D のデータを記憶領域 225C のデータと同じにするよう、記憶装置 272 又は 282 へ指示を出す。

【0106】

具体的な方法としては、記憶領域 225C から記憶領域 225D に対して非同期リモートコピーを実行して、ペア状態が Duplex になるまで待つ方法、同期リモートコピーを行う方法又は記憶領域 225C に対して更新がなされていない場合には、通常の I/O コマンドを利用してコピーを行う方法等がある。また、記憶領域 225C と記憶領域 225D との間の差分情報が利用できるのであれば、これを利用してもよい（ステップ 1404）。

【0107】

次にホスト 281 は、記憶領域 225D から記憶領域 225C に対する非同期リモートコピーを記憶装置 282 へ指示し、双方のデータ内容が Duplex になるまで待つ。なお、上記のステップで既に記憶領域 225C と記憶領域 225D のデータ内容が同一であることを利用して、非同期リモートコピーの初期化コピーを省略してもよい（ステップ 1405）。

ホスト 281 にてアプリケーション 201 を起動し、記憶領域 D227 を用いて処理を再開する。なお、再開に際して、アプリケーション 201 による復旧処理を行っても良い（ステップ 1406）。

上述の処理により、より Primary サイトでの最新の更新データが格納されている Sync サイトからデータを Async サイトへコピーした上で、Async サイトでアプリケーションの実行を再開することができる。

【0108】

図 15 は、Async サイト 180 でアプリケーションを再開した後に、通常運用状態（Primary サイトでのアプリケーション実行）へ復帰する手順例を示す図である。

まず、Primary サイト 100 を復旧させる。この際、記憶領域 225A から記憶領域 225D に関して記憶装置 120 にはリモートコピーのペア情報が存在せず、ホスト 110 には記憶領域 225A のミラー設定情報 213 の設定はされていないものとする（ステップ 1501）。

【0109】

次に、ホスト 281 は、もし存在するのであれば、記憶領域 225D から記憶

領域 225C に対する非同期リモートコピーのペア状態の情報を削除するよう、記憶装置 282 に指示する（ステップ 1502）。

次に、ホスト 281 が記憶領域 225D から記憶領域 225B に対する非同期リモートコピーの開始を記憶装置 282 に指示し、ペア状態が Duplex になるまで待つ（ステップ 1503、1504）。

ホスト 281 上のアプリケーション 201 を停止させる（ステップ 1505）。

【0110】

その後、ホスト 110 は、ステップ 1503、1504 で作成されたペアの正副関係を入れ替え、記憶領域 225B から記憶領域 225D に対して非同期リモートコピーを行うように記憶装置 120 に指示する。もし、非同期リモートコピープログラム 223 の実行でこのような処理が出来ない場合、ホスト 110 は、ペア状態を Split にした後に Simplex に遷移し、逆方向にペアを設定することを記憶装置 120 に指示する。その場合、記憶領域 225B と記憶領域 225D は同じデータを有するため、逆方向の同期リモートコピーの初期化コピーは省略されてもよい（ステップ 1506）。

【0111】

ホスト 110 には記憶領域 225B を用いるための設定が存在しないため、ホスト 110 は、記憶領域 225B を含めたミラーグループを仮想記憶領域としてアプリケーション 201 に提供する設定を行う（ステップ 1507）。

ホスト 100 上のアプリケーション 201 を再起動し、記憶領域 225B を用いた処理を再開する（ステップ 1508）。

次に、ホスト 110 はステップ 1507 のミラーグループに記憶領域 225A を追加する（ステップ 1509）。

【0112】

記憶装置 120 が、ホスト 110 の指示に基づいて、記憶領域 225A から記憶領域 225C に対する同期リモートコピーを開始する（ステップ 1510）。

その後、ホスト 110 は、記憶領域 225A のデータが記憶領域 225B のデータと一致し、ステップ 1510 のペア状態が Duplex になるまで待つ（ス

テップ1511、1512)。

【0113】

尚、ステップ1505のアプリケーション201の停止はこのステップでの処理より前に実行されてもよく、ステップ1508のアプリケーション201の再起動はそのステップよりも後に実行されてもよい。

【0114】

図16は、A s y n c サイト180でアプリケーションを再開した後に、通常運用状態 (P r i m a r y サイトでのアプリケーション実行) へ復帰する別の手順例を示す図である。

ホスト281は、アプリケーション201を停止する (ステップ1601)。

【0115】

ホスト281は、記憶領域225Cと記憶領域225Dのデータを同じにしてから、ペア状態をS i m p l e x に遷移させるように、記憶装置282に指示する (ステップ1602、1603)。

ホスト281は、記憶領域225Dから記憶領域225Aにデータをコピーするよう、記憶装置282に指示する (ステップ1604)。

その後、記憶装置120は、ホスト110の指示に基づいて、記憶領域225Aから記憶領域225Cに対して同期リモートコピーを開始する。なお、記憶領域225Aと記憶領域225Cは同じデータを有するため、初期化リモートコピーは省略されてもよい (ステップ1605)。

【0116】

ホスト110には記憶領域225Aを用いるための設定が存在しないため、ホスト110は、記憶領域225Aを含めたミラーグループを仮想記憶領域としてアプリケーション201に提供する設定を行う (ステップ1606)。

ホスト110のアプリケーション201を再起動し、記憶領域225Aを用いた処理を再開する (ステップ1607)。

ホスト110は、ステップ1606のミラーグループに記憶領域225Bを追加する (ステップ1608)。

記憶装置120は、ホスト110の指示に基づいて、記憶領域225Bから記

憶領域 225D に対する同期リモートコピーを開始する（ステップ 1609）。

【0117】

記憶領域 225B のデータが記憶領域 225A のデータ内容と一致し、ステップ 1609 のペア状態が Duplex になるまで待つ（ステップ 1610、1611）。

【0118】

尚、ステップ 1607 のアプリケーションの再起動をステップ 1609 の後に実行することで、ステップ 1609 の初期化コピーを省略することができる。その場合はさらに、ステップ 1604 で一度記憶領域 225B のコピーを同時に行い、ステップ 1608 の追加時の初期化コピーを省略してもよい。

【0119】

次に、通常運用状態において、記憶領域 225A が障害等の発生により使用不可能になった場合の処理について説明する。

図 7 は、通常運用状態において記憶領域 225A が障害等の発生により使用不可能になった場合の処理の概要を示す図である。

【0120】

記憶領域 225A が使用不能になった場合、記憶装置 120 による記憶領域 225C に対する更新データの反映は停止される。しかし、障害保護設定がされていない場合、ホスト 110 は、記憶領域 225B に対して書き込みが正常終了すればアプリケーション 201 に対しても正常終了のメッセージを通知する。このため、記憶装置 272 と記憶装置 282 との間でデータの一致が取れない、具体的には、記憶装置 272 には存在しないデータが記憶装置 120 及び 282 に存在することになる。

【0121】

上述の不都合を解消して、記憶装置 272 におけるデータ消失の防止を優先させるために、上述の事態が発生した場合、本実施形態においては、記憶領域 225B から記憶領域 225C へデータを反映させる設定に変更する手順を実行する。

図 17 は、記憶領域 225A の障害の発見から上記設定を変更して同期リモ

トコピーを再開する手順例を示す図である。

【0122】

ホスト110が、アプリケーション201での書き込み要求に基づいて、記憶領域225Aと記憶領域225Bへ書き込み要求を発行する（ステップ1701、1702）。

本例においては、記憶領域225Aに障害が発生しているので、記憶装置120は、記憶領域225Aに対する書き込み失敗をホスト110に報告する（ステップ1703）。

【0123】

書き込み失敗を受信したホスト110は、記憶領域225Aと記憶領域225Cとのペア状態の情報を削除するように、記憶装置120に指示する（ステップ1704）。

又、ホスト110は、記憶領域225Bと記憶領域225Dとのペア状態をSimplexに遷移させるように、記憶装置120に指示する（ステップ1705）。

【0124】

記憶装置120は、ホスト110の指示に基づいて、記憶領域225Bから記憶領域225Cに対して同期リモートコピーを開始する。なお、記憶領域225Bにだけ更新されたデータが存在するので、記憶領域225Bと記憶領域225Cとのデータの内容は、常に同じとは限らない。しかし、この場合でも以下に示す手順で初期化コピーを省略できる（ステップ1706）。

【0125】

ステップ1706にて初期化コピーが省略された場合、ホスト110が、ステップ1701にて行われた記憶領域225Bに対する書き込み要求を記憶領域225Bに対して再度発行する。この処理で、記憶領域225Bと記憶領域225Cのデータの内容が同じになる。なお、書き込み要求の再発行は、ミラープログラム212の実行に基づいてホスト110が行っても良いし、アプリケーション201の実行に基づいてホスト110が行っても良いし、それ以外のソフトウェアの実行に基づいてホスト110が行っても良い（ステップ1707）。

【0126】

なお、上述の処理のステップ1704を実行中にSyncサイト170に障害が発生した場合、記憶領域225Dの一部に記憶領域225Bのデータが反映された状態となり、Asyncサイトにおいてアプリケーション201の再開ができなくなる場合がある。これに対する対策としては、ステップ1704を実行する前にAsyncサイトで記憶領域225Dのバックアップを行うことが考えられる。

【0127】

なお、本実施形態では、説明の簡略化のために、アプリケーション201はミラープログラム212が提供する一つの記憶領域のみを用いるように記載していたが、本発明はこれに制限されない。具体的には、アプリケーション201が二つ以上の記憶領域を用いる場合、上述した記憶領域225Aから記憶領域225Dの組を複数個用意し、ミラーグループも複数個作成する。また、図9に示した書き込み処理以外は、記憶領域が関係する処理を複数回繰り返すことで対応できる。

【0128】

図9に示した処理の場合も同様に、処理を記憶領域の組の数分繰り返すことが考えられる。しかし、複数個存在する記憶領域225Aのいずれかが障害Suspendによる書き込み禁止がなされた場合にそれ以外の記憶領域225Aに対する書き込みも拒否することで、記憶装置120は、記憶領域225A全体としてみた場合の一貫性を維持できる。したがってこの場合、図9のステップ907やステップ912の処理で扱われる記憶領域の数を一つのミラーグループに所属する記憶領域の数とするのではなく、それぞれ複数個存在する記憶領域225Aと記憶領域225Bの組の数と考える。これにより、全ての記憶領域について処理を繰り返すよりも処理ステップ数を削減することが可能となる。

【0129】

また、本実施形態において、記憶領域225Aと記憶領域225Bは別々な記憶装置内に存在してもよい。この場合、ミラープログラム212は、データの複製を各々異なる記憶装置へ転送する。さらに、ミラープログラム212は記憶装

置 120 に存在してもよい。この場合、システム（特に記憶装置 120）は、以下に示す手順に従って、Sync サイト 170 と Async サイト 180 にリモートコピーを行う。

【0130】

- (1) ホスト 110 から記憶装置 120 に対して書き込み要求が送られる。
- (2) 記憶装置 120 内のミラープログラム 212 が、書き込み要求に従って記憶領域 225 A と記憶領域 225 B に書き込みを行う。
- (3) 同期リモートコピープログラム 222 は記憶領域 225 A に対する書き込みを受け取り、Sync サイトの記憶領域 225 C に対して書き込み要求を転送したことを確認してからミラープログラム 212 に書き込み完了を返す。

【0131】

- (4) 一方、非同期リモートコピープログラム 223 は記憶領域 225 B に対する書き込みをログエントリ化してからミラープログラム 212 に書き込み完了を返す。
- (5) ミラープログラム 212 は両方の書き込み完了が返されたらホスト 110 に対して書き込み完了を返す。
- (6) 非同期リモートコピープログラム 223 はログエントリを Async サイトに転送し、順序関係を保って記憶領域 225 D に書き込みを反映させる。

【0132】

以上の動作中に Primary サイト 100 に障害が発生した場合は、システムは Sync サイトか Async サイトのどちらかでアプリケーションを再開する。

尚、この場合のミラーグループの設定等は、ホスト 110 に設定用のアプリケーションを導入して管理者がそのアプリケーションを使用して設定するか、記憶装置 110 が有する管理端末から管理者が設定する等が考えられる。

【0133】

次に、本発明の第 2 の実施形態について説明する。

図 18 は、第 2 の実施形態であるシステム概要を示す図である。本実施形態と第 1 の実施形態との相違点は、各サイトに中間サーバー 1801、1802、1

803が追加されたことである。中間サーバーとは、各サイトのホストと記憶装置との間のI/Oを中継する働きを持つ計算機で、例えば、NFSやCIFSのファイルサーバーや、仮想的なボリュームを提供するバーチャリゼーションサーバーなどが考えられる。

【0134】

なお、中間サーバーは記憶装置120内や、ホスト110内に存在してもよい（例えば、機能が豊富なネットワークインターフェースとして）。特に、Primaryサイト100の中間サーバー1801にはミラープログラム212が格納され、第1の実施形態においてホスト110が行っていたミラーリングの処理を中間サーバー1801が実行する。これにより、ホスト110に何ら変更を加えることなく、第1の実施形態で説明した処理が実施可能となる。

【0135】

なお、各サイトのホスト、中間サーバー、記憶装置を接続するネットワークは第一の実施例同様、どのような転送媒体やトポロジーであってもよい。

【0136】

次に、第1の実施形態を応用した第3の実施形態について説明する。

図19は、第3の実施形態のシステムの概要を示すブロック図である。なお、本図においては以下説明に必要な部分のみ図示して残りを省略したが、本実施形態でも第1の実施形態における各種プログラムやハードウェアは存在する。ただし、非同期リモートコピープログラム223は本実施形態では必須ではない。

【0137】

第1と第2の実施形態では記憶装置に非同期リモートコピープログラムが存在することを前提としたが、本実施形態では、データベースシステムで生成されるログの転送によって、非同期リモートコピーの代替とする。以後、このようなリモートコピーをデータベースによる非同期リモートコピーと呼ぶ。

【0138】

データベース1907は、アプリケーション201がクエリーの処理を要求するデータベースを制御するプログラムである。データベース1907を実行中のホスト110は、アプリケーション201が要求したクエリーを完了（以下「コ

ミット」) する際に、必ず記憶装置 120 が有するログ用記憶領域 AL 1902 へログを書き込む。また、図示はしていないが、ホスト 110 は、データベース 1907 の記憶領域 AD 1901 (テーブルを保存するための記憶領域) 用のバッファを持ち、ログに対応した更新をすぐには書き込まず、バックグラウンドで記憶領域 AD 1901 へ書き込む。

【0139】

また、データベース 1907 では、コミットされ、かつ更新内容が既に記憶領域 AL 1902 に書き込まれたクエリーの最新時刻、具体的には書き込まれたログの最新時刻 (またはそれより少し古い時刻) がログとともに管理されている。従って、ホスト 110 は、データベースの復旧の際にはこの時刻のログから復旧処理を開始することになる。また、ログにはクエリーによって変更されるデータの変更前と変更後のデータが登録される。

【0140】

なお、データベースでは、前述した時間の代わりに、前述の時間をクエリー毎に割り当てるなどしたシーケンシャルな ID を用いたり、これらの両方を情報として持つことがある。ここで、ID の場合は予め連続した番号を割り振るというルールを設定しておけば欠番が認識できるという特長を持つが、それ以外はどちらを使用しても特に差異がないので、以後の説明では時間を使用した場合について説明する。

【0141】

記憶領域 CD 1093 は、記憶領域 AD 1901 と同期リモートコピーのペアとなる記憶領域である。記憶領域 CL 1094 は、記憶領域 AL 1902 と同期リモートコピーのペアとなる記憶領域である。

【0142】

ホスト 110 は、DB ログ送信プログラム 1908 を実行することで、データベース 1907 の処理で生成されるログを Async サイト 180 のホスト 281 へ送信する。なお、ログ転送の際、ホスト 110 がログをログ用記憶領域 AL 1902 から定期的に取り出してホスト 281 へ送付するが、これ以外に、ホスト 110 がログ用記憶領域 AL 1902 にデータを保存する前に、直接データを

ホスト 281 へ転送しても良い。

【0143】

ホスト 281 は、DB ログ受信プログラム 1909 を実行することで、ホスト 110 が転送したログを受け取り、ログ用記憶領域 DL 1906 へログの追加を行い、記憶領域 DD 1905 へ更新データの書き込みを行う。既に述べたように、ログにはクエリーによって変更されたデータのデータの変更前と変更後のデータの片方または両方が存在するため、Async サイト 180 がこのログを元に記憶領域 D 1905 に対して書き込みを行えば、非同期リモートコピーと同じ処理を行うことができる。

【0144】

なお、転送されるログの代わりにアプリケーション 201 で生成されたクエリーそのものを転送し、このクエリーをホスト 281 上で再実行する方法も考えられる。

【0145】

なお、本実施形態においても、第 1 の実施形態で説明した図 10 から図 16 の処理と同様の処理が行われる。ただし、各処理において、記憶領域 225 A を記憶領域 AD 1901 及びログ用記憶領域 AL 1092 に、記憶領域 225 C を記憶領域 CD 1903 及びログ用記憶領域 CL 1094 に、記憶領域 225 D を記憶領域 DD 1905 及びログ用記憶領域 DL 1096 に入れ替える。又、図 10 から図 16 の各処理において、ミラーグループに関連する処理及び記憶領域 225 A と記憶領域 225 B との間のデータコピー処理を削除し、これ以外の記憶領域 225 B が関係する処理を記憶領域 225 A と読み替える。

【0146】

また、データベース 1907 は各々役割が異なる二つの記憶領域を用いるように記述したが、一つの記憶領域をそれに割り当てても良く、また 3 つ以上の記憶領域を用いても良い。

【0147】

なお、Primary サイトが障害停止した際に、Sync サイトと Async サイト間で非同期のリモートコピーを作成することでアプリケーションの実行

を再開する場合については、以下に示す方法を用いることで、図10で説明した処理手順よりも初期化コピーの処理時間を短縮できる。

【0148】

図20は、Primaryサイトが時刻Time0にて障害停止した場合のSyncサイト170とAsyncサイト180の状態を示す図である。

まず、ログ用記憶領域CL1904は同期リモートコピー222によって常に最新の状態となっていたため、以下の関係が成り立つ。なお以下で、ログ記憶領域CL1904が有する最新のログの時刻をTimeCLNew、記憶領域CD1903に反映され、不必要となったログの最新（またはそれより少し古い）時刻をTimeCDNew、ログ記憶領域DL1906が有する最新のログの時刻をTimeDLNew、記憶領域DD1905に反映され、不必要となったログの最新（またはそれより少し古い）時刻をTimeDDNewとする。

【0149】

- (A) TimeCLNew以後に実行完了したクエリーは存在しない
- (B) TimeCLNewはTimeCDNewと同じ時刻か、より新しい時刻（データベースはログ用の記憶領域に先に書き込むため）である。
- (C) TimeDLNewはTimeDDNewと同じ時刻か、より新しい時刻（データベースはログ用の記憶領域に先に書き込むため）である。
- (D) TimeCLNewはTimeDLNewと同じ時刻か、より新しい時刻（Primaryサイト100からAsyncサイト180へのコピーは非同期で行われるため）である。
- (E) TimeCDNewはTimeDDNewと同じ時刻か、より新しい時刻（Primaryサイト100からAsyncサイト180へのコピーは非同期で行われるため）である。

【0150】

図21は、図20の状態において、Syncサイト170からAsyncサイト180へデータベースによる非同期リモートコピーを行うための手順を示す図である。

ホスト271及びホスト281は、TimeCLNew、TimeCDNew

、TimeDLNew、及びTimeDDNewの情報を収集する（ステップ2101）。

【0151】

ホスト271及びホスト281は、ログ用記憶領域CL1904が有するログの最古の時刻TimeCLOldと、ログ用記憶領域DL1904が有するログの最古の時刻TimeDLOldの情報を収集する（ステップ2102）。

次に、ホスト271又はホスト281は、TimeCLOldとTimeDLNewとを比較する（ステップ2103）。

【0152】

TimeCLOldがTimeDLNewより一つだけ新しい時刻、具体的には、先ほど述べた更新順序を示すカウンタで一つ分新しい時刻と同じか、より古い時刻ならば、Syncサイト170は、TimeDLNewより一つだけ新しい時刻からTimeCLNewまでのログをAsyncサイト180へ転送する。これによって、Syncサイト170に存在するログはすべてAsyncサイト180にも存在するようになる（ステップ2104）。

【0153】

一方、TimeCLOldがTimeDLNewより二つ以上カウンタの値が新しい時刻である場合、Asyncサイト180に転送すべきログの一部をSyncサイトですでに削除してしまったことを意味するため、記憶装置272は、ログ用記憶領域CL1904をログ用記憶領域DL1906へ、記憶領域CD1903を記憶領域DD1905へコピーする（ステップ2109）。

【0154】

ステップ2104又はステップ2109の終了後、Syncサイト170にてデータベース1907の復旧処理を行う。通常、データベースではログを用いて復旧処理が行われる。具体的には、ログ用記憶領域に蓄積されたログを古い順からデータベースのデータに適用し、最後までログを適用した後に、コミットがされていないクエリーに関係したログについてはロールバックが行われる（ステップ2105）。

【0155】

A s y n c サイトにてステップ 2105 と同様の復旧処理を実施する（ステップ 2106）。

ホスト 271 又はホスト 281 が、S y n c サイト 170 から A s y n c サイト 180 へのデータベースによる非同期リモートコピーを開始する（ステップ 107）。

データベース 1907 はアプリケーション 201 と共に処理を再開する（ステップ 2108）。

【0156】

なお、コピー量削減のため、ログ用記憶領域間のデータコピーについては、ログ用記憶領域 C L 1904 とログ用記憶領域 D L 1906 が有するログを比較して、ログ用記憶領域 D L 1906 が有していないログ（かつ復旧処理に必要な期間）だけを転送する方法が考えられる。また、記憶領域に保存されているテーブルを構成する各ページにはそのページに適用された最新のログを示す情報が含まれることがあり、サイト間でこれを比較し、差異のあるページだけをコピーすることも考えられる。

【0157】

また、S y n c サイト 170 から A s y n c サイト 180 へ最新データをコピーしてから A s y n c サイトで復旧する場合は、ステップ 2106 までの処理はそのまま実行し、それ以降の処理で S y n c サイト 170 と A s y n c サイト 180 の関係を入れ替えることで対処可能となる。

【0158】

次に、本発明の第 4 の実施形態について述べる。本実施形態では、同期リモートコピーは他の実施形態と同様に記憶装置で行い、非同期リモートコピーはホスト上のソフトウェアによって実現される。さらに、本実施形態では、ホスト 110 はジャーナルファイルシステム 2202 を用いることで、P r i m a r y サイトが障害停止した際の復旧処理にて行われる記憶領域 225C と記憶領域 225D 間の非同期リモートコピーの初期化コピーにかかる時間の短縮化を行う。

【0159】

図 22 は、本実施形態の構成概要を示したブロック図である。

ホスト 110 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行することで、ファイルシステム 2202 やアプリケーション 201 からの書き込み要求を受付け、書き込みデータと時間順序情報をまとめてエントリとして保持した後に記憶領域 225A に対して書き込み処理を行う。又、ホスト 110 は、同プログラム 2201 を実行することで、作成したエントリをホスト 281 へ送受信したり、ホスト 281 の同プログラムについては受信したエントリから記憶領域に書き込みデータを反映させる処理も行う。

【0160】

ジャーナルファイルシステム 2202 とは、ファイルシステムのファイル毎に存在するメタデータの変更をログの形で記憶領域上に保持するファイルシステムのことで、ホスト障害時のファイルシステムのメタデータチェックの時間短縮を実現したファイルシステムである。なお、メタデータのログを蓄積する領域（メタデータログ領域）は記憶領域の定められた範囲に割り当てることが考えられるが、メタデータログ領域は拡張可能であってもよい。いずれの場合でも、メタデータ領域に対して書き込みが行われれば同期リモートコピーによって同様の書き込みデータが記憶領域 225C にも存在することになる。

【0161】

さらにファイルが更新される場合、メタデータの更新契機は、ファイルオープン、クローズ、キャッシングされたブロックデータの更新、及び記憶領域上のブロックデータの更新等が考えられる。しかし、更新対象ファイルのオープンからクローズまでの間に一回でもメタデータの更新が行われれば、上記の条件が発生した場合に必ずメタデータを更新する処理が行われる必要はなく、これ以外の契機で更新が行われてもよい。また、メタデータログ領域以外の記憶領域上のメタデータの更新はメタデータログ領域の更新と同時であってもよく、非同期であってもよい。

【0162】

以下、本実施形態で、ファイルの更新処理がどのように行われるかについて説明する。

(1) アプリケーションがファイルを書き込み権限つきでオープンする。

(2) データを更新する。ジャーナルファイルシステム 2202 はバッファキャッシュを経由して、ブロック書き込みの要求を非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 に送信する。ホスト 110 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行して書き込み要求を記憶領域 225A に対して発行し、エントリを作成してホスト 110 で保持する。その後、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 はジャーナルファイルシステムに対して書き込み完了メッセージを送信する。

【0163】

(3) 記憶装置 120 は、同期リモートコピープログラム 222 を実行して、記憶領域 225A に書き込まれた更新データを記憶領域 225C に対して反映させる。

(4) ホスト 110 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行して、(2) にて作成したエントリを Async サイト 180 へ転送し、転送が完了したら当該エントリを削除する。

(5) Async サイト 180 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行して、受け取ったエントリを元に順序を守りながら記憶領域 225D に更新データを反映させる。

【0164】

(6) アプリケーション 201 はファイルをクローズする。この段階で最低一回はメタデータの更新が行われている。その際の更新手順はメタデータログ領域内のメタデータとそれ以外のメタデータの両方について以下の (A) から (D) のステップをたどり、Sync サイト 170 と Async サイト 180 へ伝達される。

【0165】

(A) ホスト 110 がメタデータを更新する。ジャーナルファイルシステム 2202 はブロック書き込みの要求を非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 に送信する。ホスト 110 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行することで、書き込み要求を

記憶領域 225A に対して発行し、エントリを作成してホスト 110 上で保持する。その後、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 は、ジャーナルファイルシステムに対して書き込み完了メッセージを送信する。

【0166】

(B) 記憶装置 120 は、同期リモートコピープログラム 222 を実行することで、記憶領域 225A に書き込まれた更新メタデータを記憶領域 225C に対して反映させる。

(C) ホスト 110 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行して、(A) にて作成したエントリを Async サイト 180 へ転送し、転送が完了したら当該エントリを削除する。

【0167】

(D) Async サイト 180 のホスト 281 は、非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 を実行して、受け取ったエントリを元に順序を守りながら記憶領域 225D に更新メタデータを反映させる。

【0168】

なお、メタデータログ領域は状況によっては容量不足となることもあるので、古いメタデータのログを削除する必要がある。本実施形態では、以下のステップにてメタデータのログを削除することで、メタデータログ領域にログが存在しないファイルは記憶領域 C 227 と記憶領域 D 228 で同じデータを有することを保障する。

(1) ジャーナルファイルシステム 2202 は削除するログを決定する。

【0169】

(2) 削除対象のログがその他のログと同じファイルに関連しているのであれば、削除対象のログを削除して処理が終わる。

(3) 削除対象のログがその他のログと同じファイルに関連していないのであれば、ジャーナルファイルシステム 2202 はこのファイルに関連する記憶領域 (メタデータについては対象としなくてもよい) のブロックアドレスを得る。

【0170】

(4) P r i m a r y サイト 100 の非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 の実行によって作成されるエントリを検索し、(3) で得たブロックアドレスを更新するエントリが P r i m a r y サイト 100 に存在しなくなるまで待つ

(5) 削除対象のログを削除する。

【0171】

本実施形態において、P r i m a r y サイト 100 の障害停止時の復旧処理で S y n c サイト 170 から A s y n c サイト 180 へ非同期リモートコピー機能つき仮想記憶領域提供プログラム 2201 による非同期リモートコピーを行うための手順は以下の通りである。

(1) S y n c サイト 170 の記憶領域 225C のメタデータログ領域に存在するログを検索し、更新が発生したファイルの一覧を得る。

【0172】

(2) ホスト 271 は、(1) にて得たファイルのデータを記憶領域 225C から記憶領域 225D に対してコピーするよう、記憶装置 272 へ指示する。

(3) S y n c サイト 170 から A s y n c サイト 180 へのデータベースによる非同期リモートコピーを開始する。

(4) S y n c サイト 170 でアプリケーション 201 の処理を再開する。

【0173】

また、本実施形態にて、S y n c サイト 170 から A s y n c サイト 180 へ最新データをコピーしてから A s y n c サイトで復旧する場合は、上述した処理のうち(3)までの処理はそのまま実行し、それ以降の処理で S y n c サイト 170 と A s y n c サイト 180 の関係を入れ替えれば良い。

【0174】

上述した全ての実施形態において、通常運用時の A s y n c サイト 180 に対する更新データの伝達は非同期リモートコピーによって継続的に行われるため、P r i m a r y サイトと S y n c サイトの両方が障害停止した場合に消失する更新データ量が少なくなる。

また、記憶装置が特殊な機能を有さずに、複数の記憶領域を用いて同じデータ

の同期リモートコピーと非同期リモートコピーを実現することが可能なため、幅広い記憶装置に対して本発明を適用できる。

【0175】

【発明の効果】

本発明によれば、PrimaryサイトとSyncサイトの両方が障害停止した場合に消失する更新データ量が少なくすることができる。

【図面の簡単な説明】

【図1】

システムのハードウェア図である。

【図2】

本実施例の機能構成を表したブロック図である。

【図3】

本実施例のシステム初期化手順を示したフロー図である。

【図4】

本実施例の通常運用状態を示したブロック図である。

【図5】

Primaryサイトが障害停止した場合にSyncサイトでアプリケーション処理を再開した場合の状態を示したブロック図である。

【図6】

ミラーグループへ記憶領域を追加する手順を示したフロー図である。

【図7】

Syncサイトに対応した記憶領域に障害が発生した状態を示したブロック図である。

【図8】

設定情報を示す図である。

【図9】

記憶領域同期保護が有効な場合のPrimaryサイトでの書き込み処理の手順について示したフロー図である。

【図10】

P r i m a r y サイトに障害が発生した場合に、S y n c サイトでアプリケーション処理を再開する手順を示したフロー図である。

【図 1 1】

S y n c サイトでアプリケーション処理を再開後に P r i m a r y サイトが復旧した際の通常運用状態への復帰手順を示したフロー図である。

【図 1 2】

S y n c サイトでアプリケーション処理を再開後に P r i m a r y サイトが復旧した際の通常運用状態への復帰手順を示したフロー図である。

【図 1 3】

P r i m a r y サイトやさらに S y n c サイトが障害停止した際に A s y n c サイトでアプリケーション処理を再開する処理手順を示したフロー図である。

【図 1 4】

P r i m a r y サイトが障害停止した際に A s y n c サイトでアプリケーション処理を再開する手順を示したフロー図である。

【図 1 5】

A s y n c サイトでアプリケーション処理を再開後に P r i m a r y サイトやさらに S y n c サイトが復旧した際の通常運用状態への復帰手順を示したフロー図である。

【図 1 6】

A s y n c サイトでアプリケーション処理を再開後に P r i m a r y サイトやさらに S y n c サイトが復旧した際の通常運用状態への復帰手順を示したフロー図である。

【図 1 7】

記憶領域 A に障害が発生した場合に、同期リモートコピーを継続するための処理手順を示したフロー図である。

【図 1 8】

第 2 の実施形態の構成例を表したブロック図である。

【図 1 9】

第 3 の実施形態の構成例を表したブロック図である。

【図 2 0】

第 3 の実施形態において、P r i m a r y サイトが障害停止した後の状態を示したブロック図である。

【図 2 1】

第 3 の実施形態における、P r i m a r y サイトが障害停止した際の S y n c サイトでのアプリケーション処理を再開する手順を示した図である。

【図 2 2】

第 4 の実施形態の構成例を表したブロック図である。

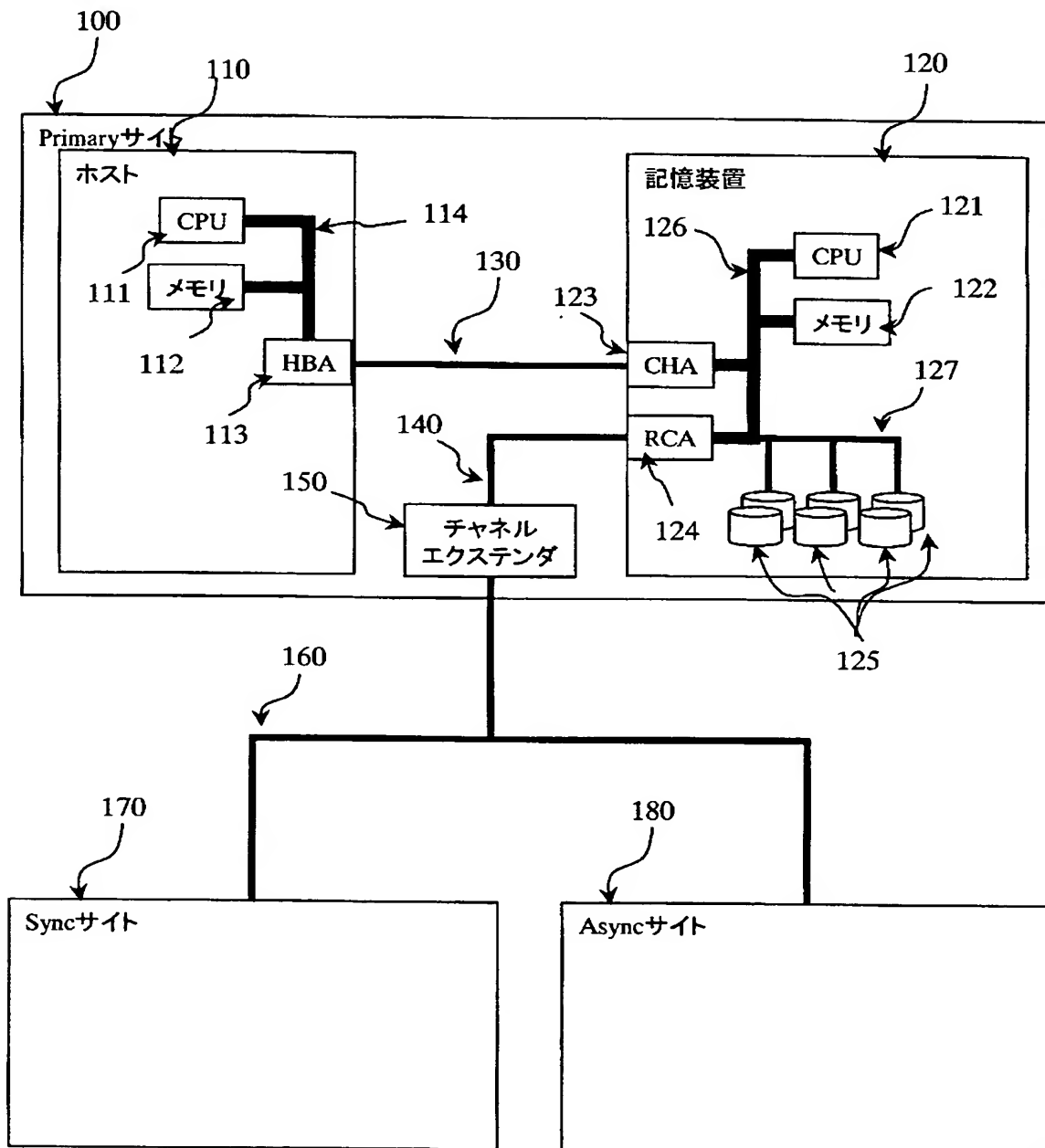
【符号の説明】

1 0 0 … P r i m a r y サイト、1 7 0 … S y n c サイト、1 8 0 … A s y n c サイト、1 1 0 … ホスト。

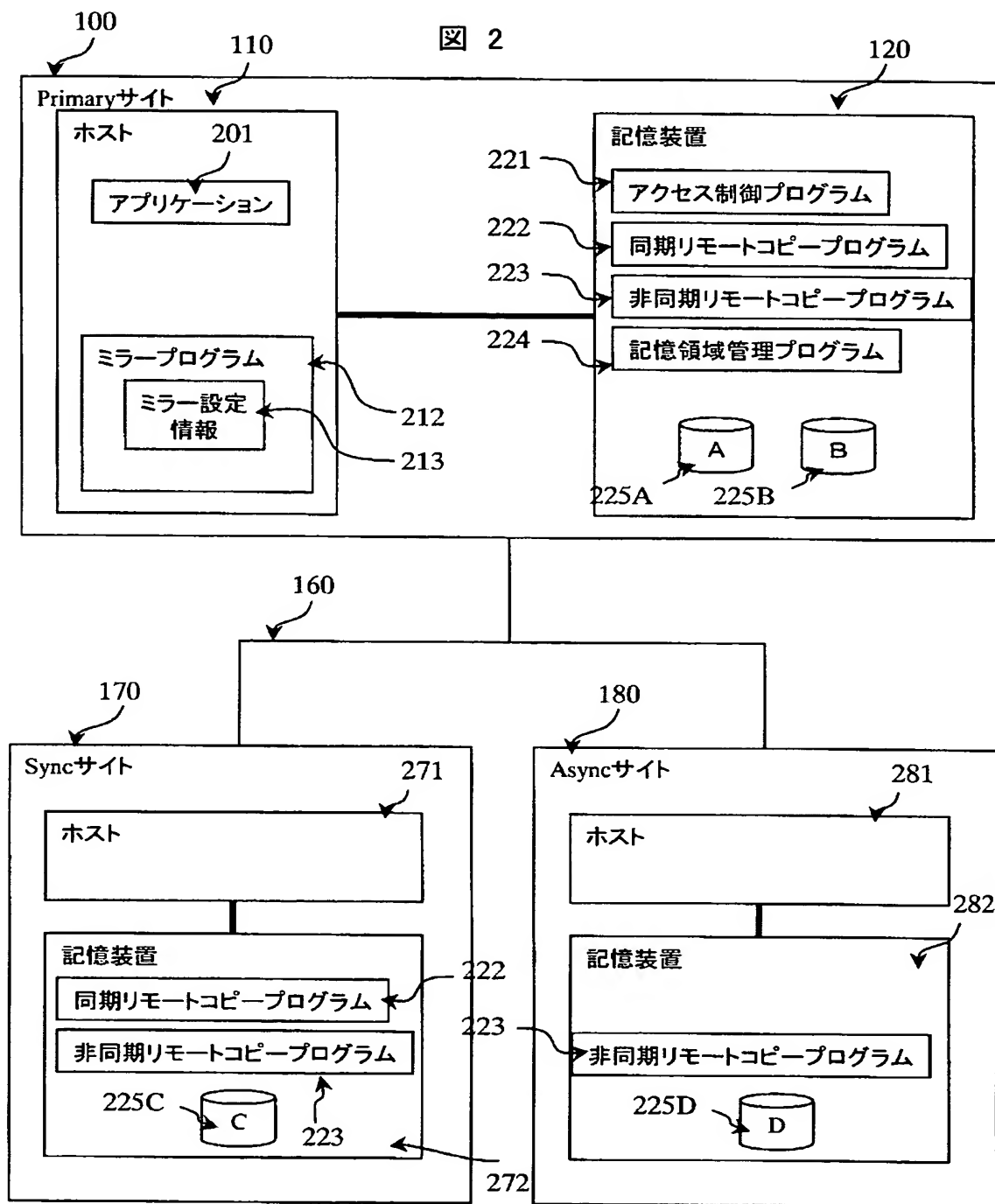
【書類名】 図面

【図 1】

図 1

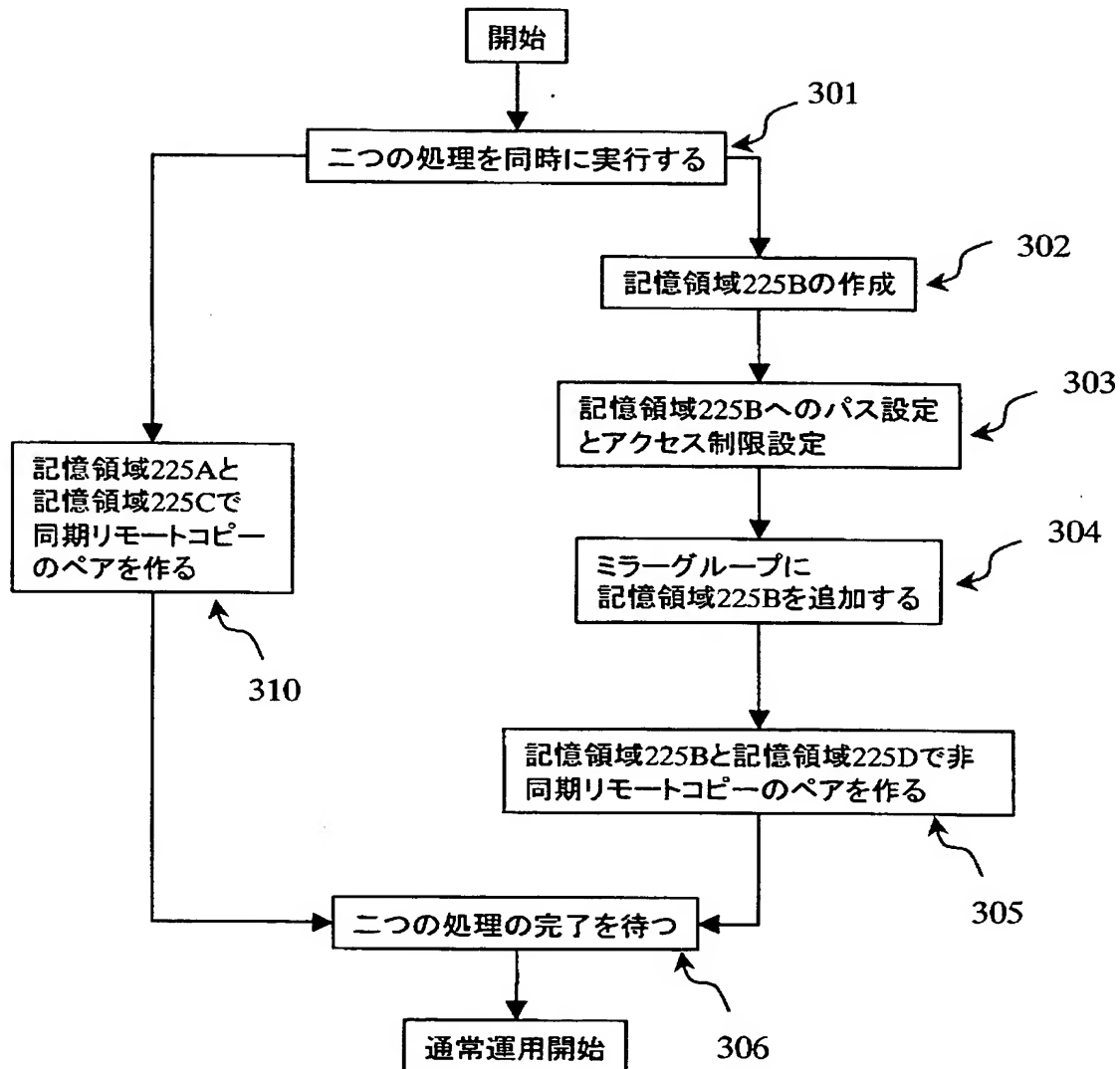


【図 2】

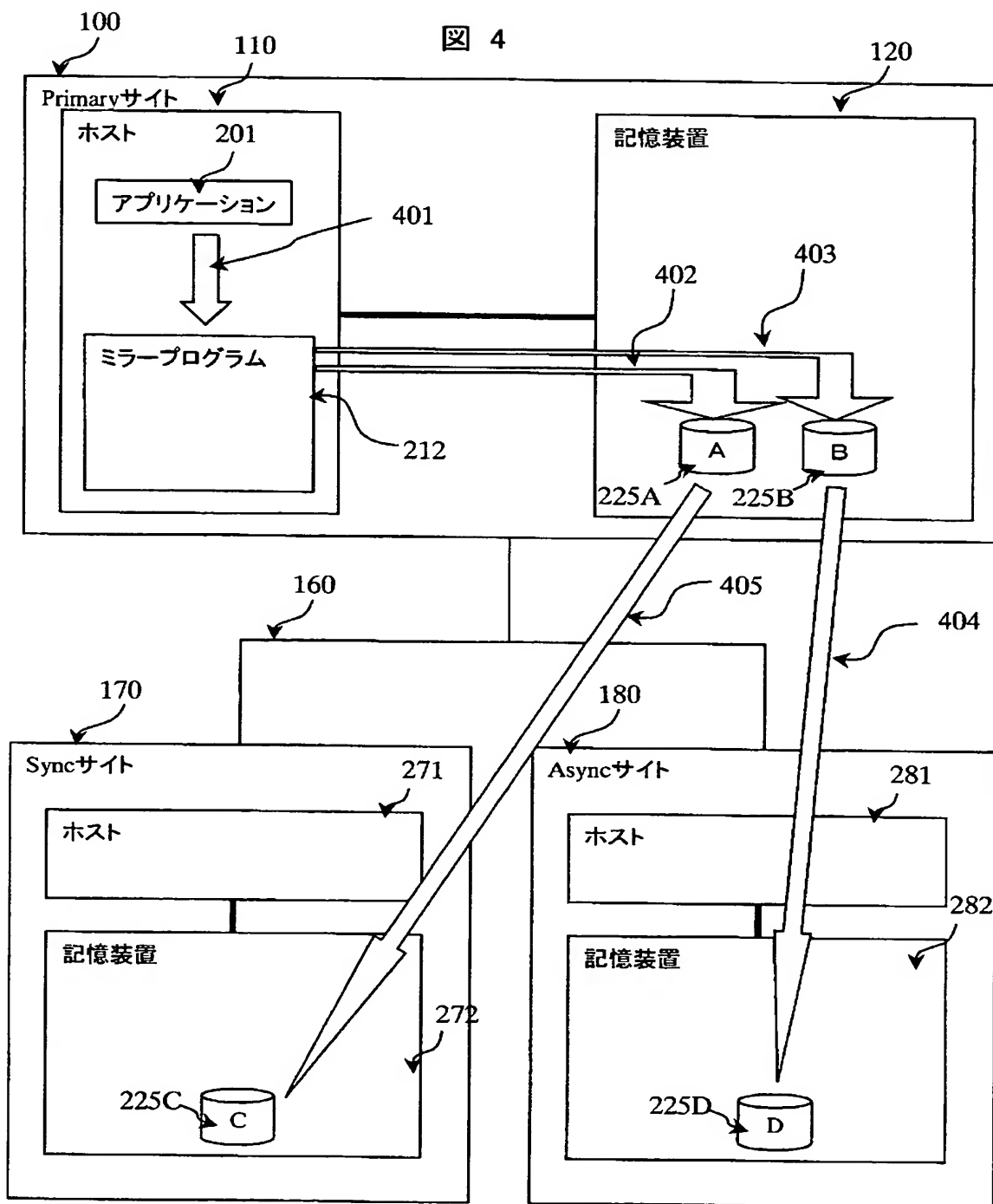


【図 3】

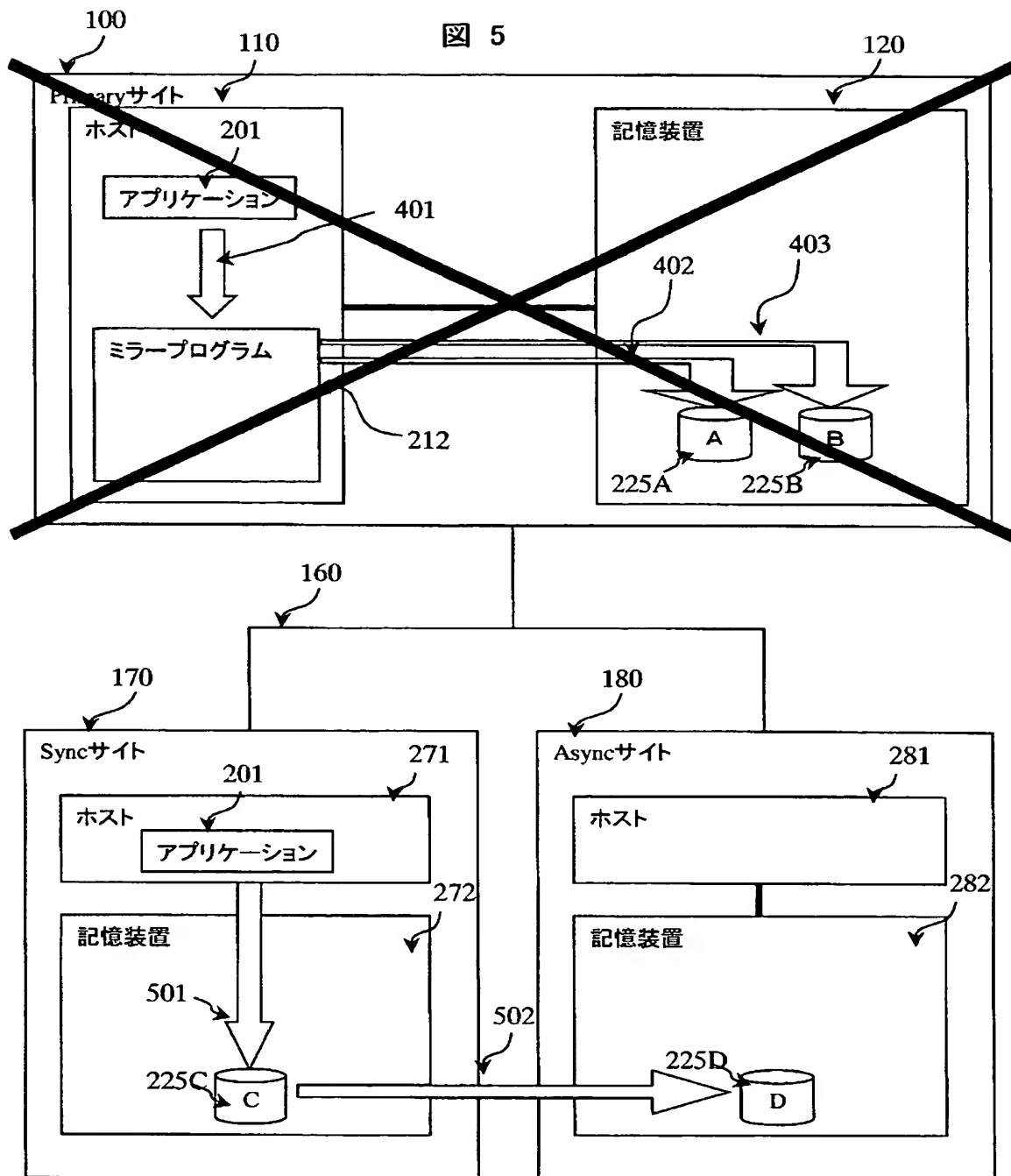
図 3



【図 4】

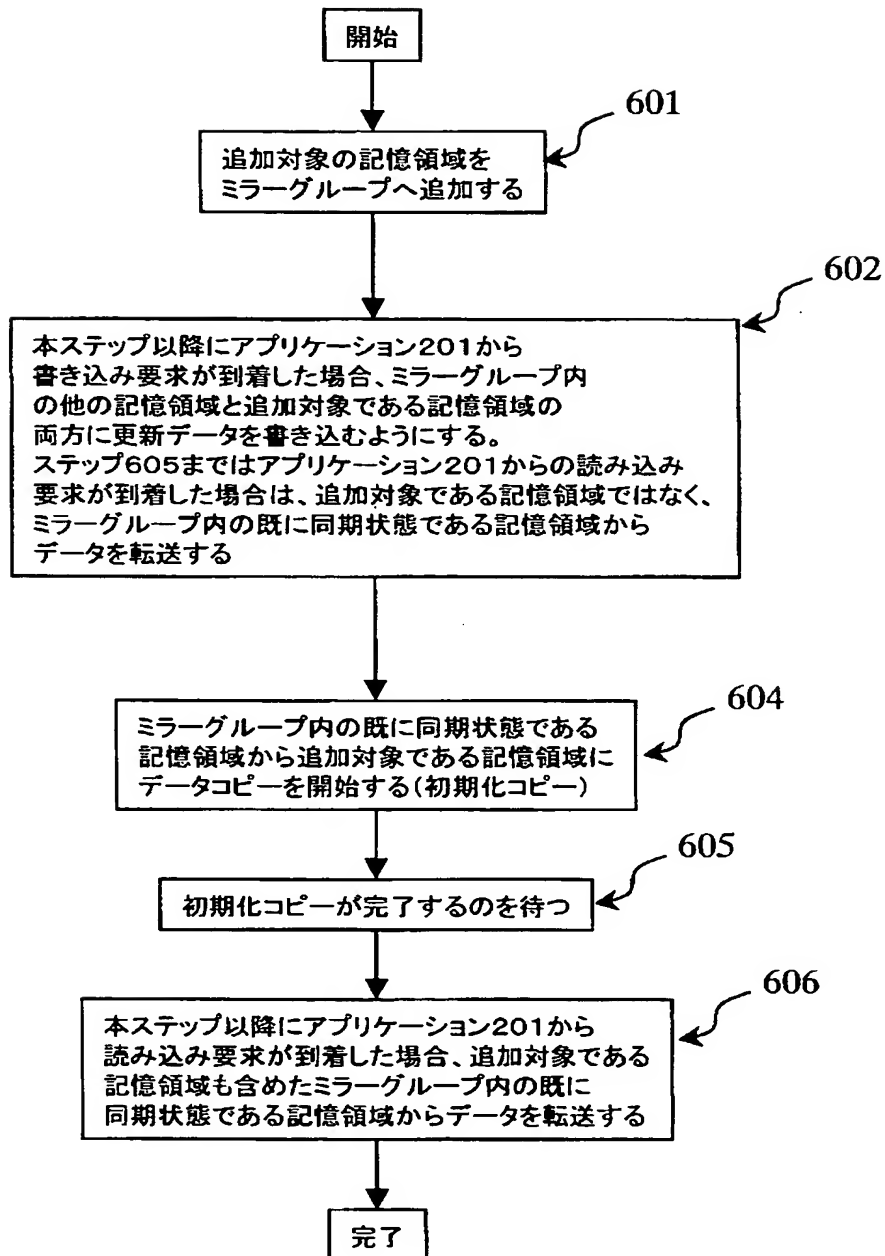


【図 5】

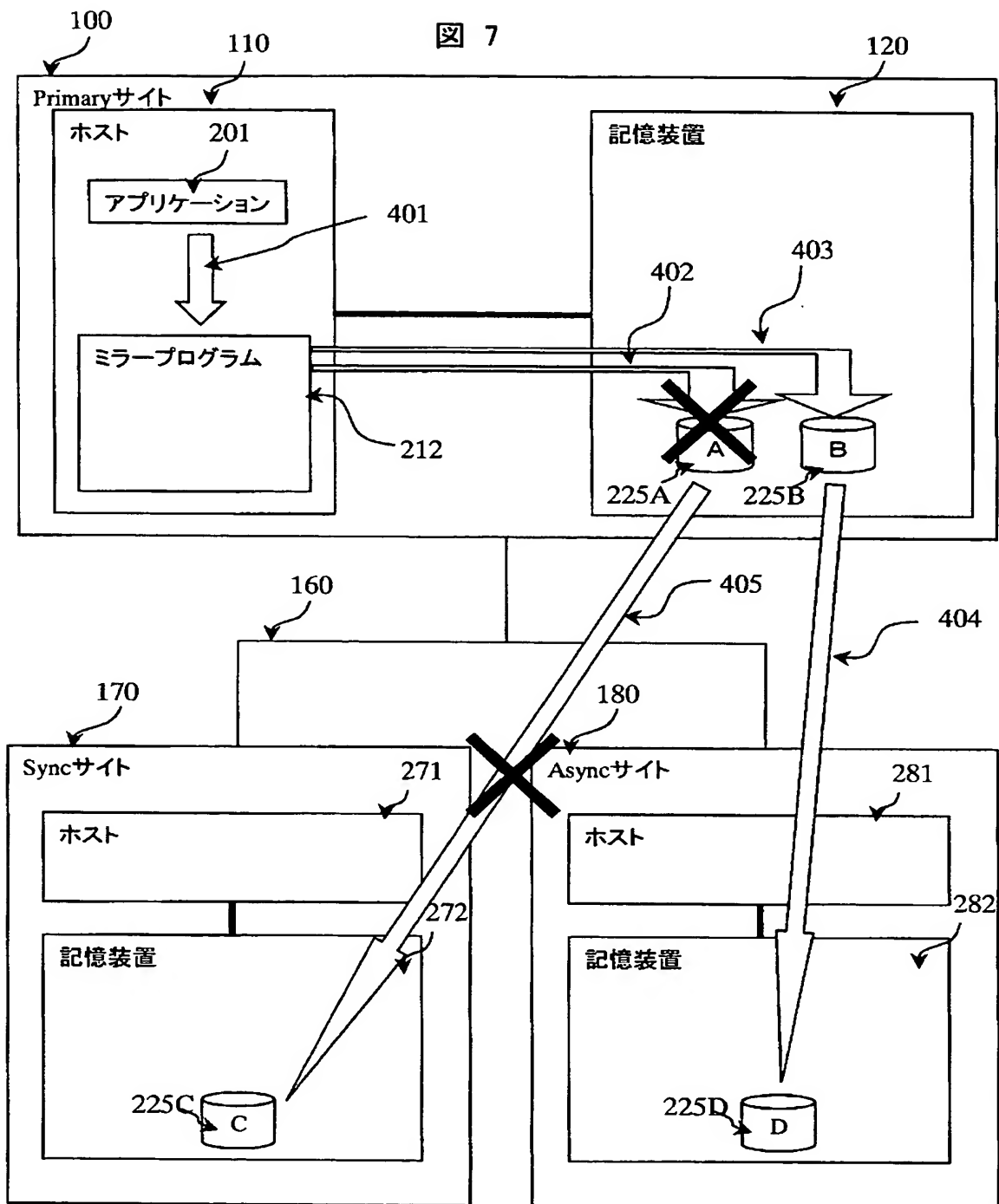


【図 6】

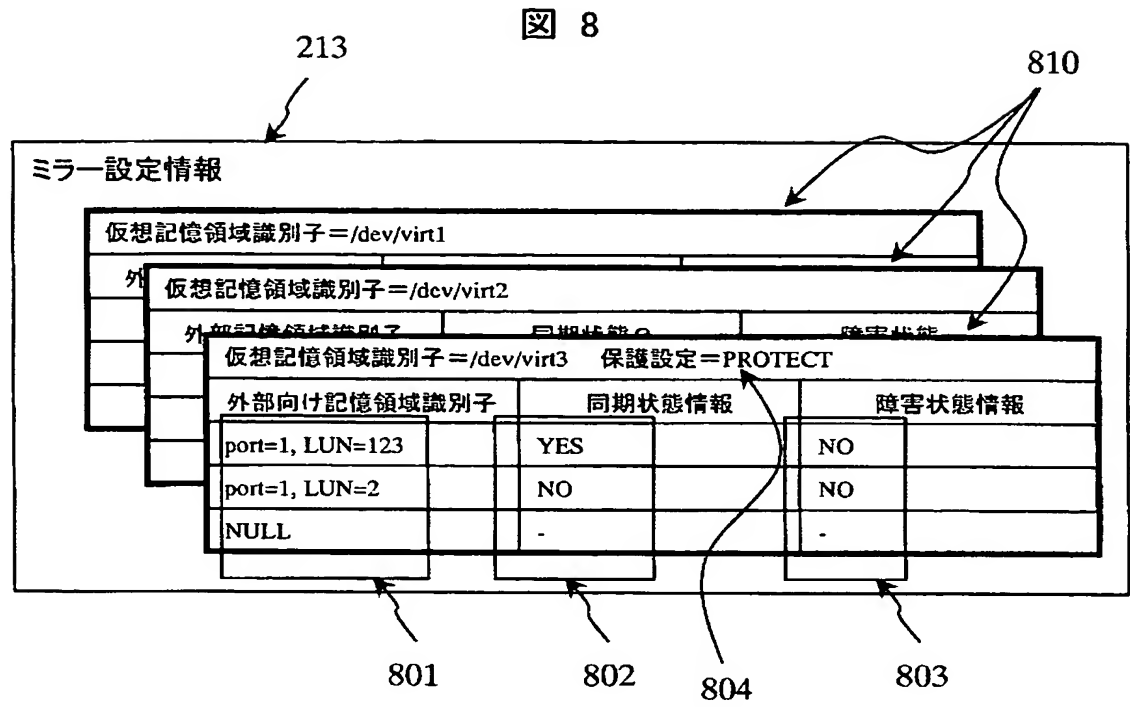
図 6



【図 7】

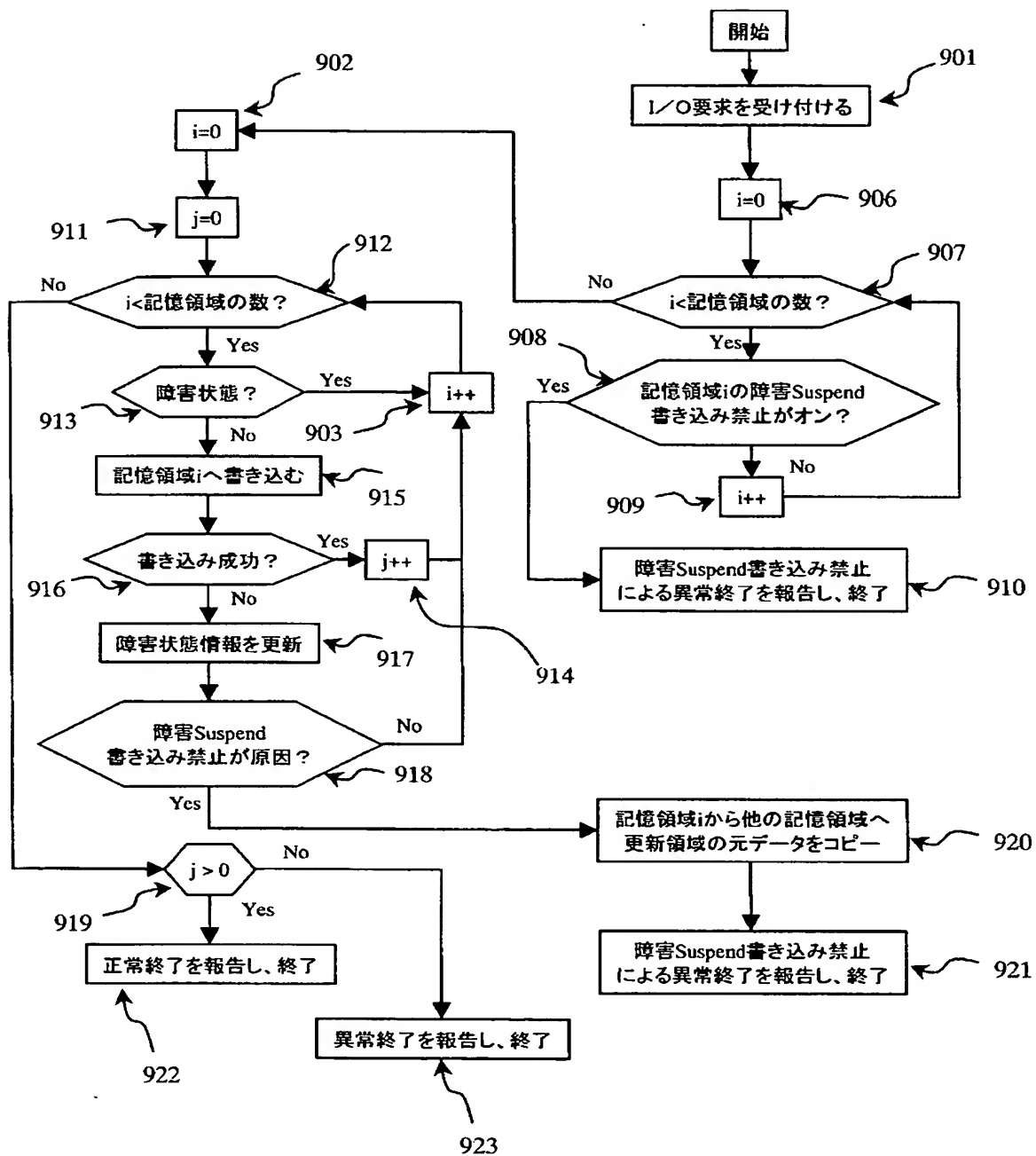


【図 8】



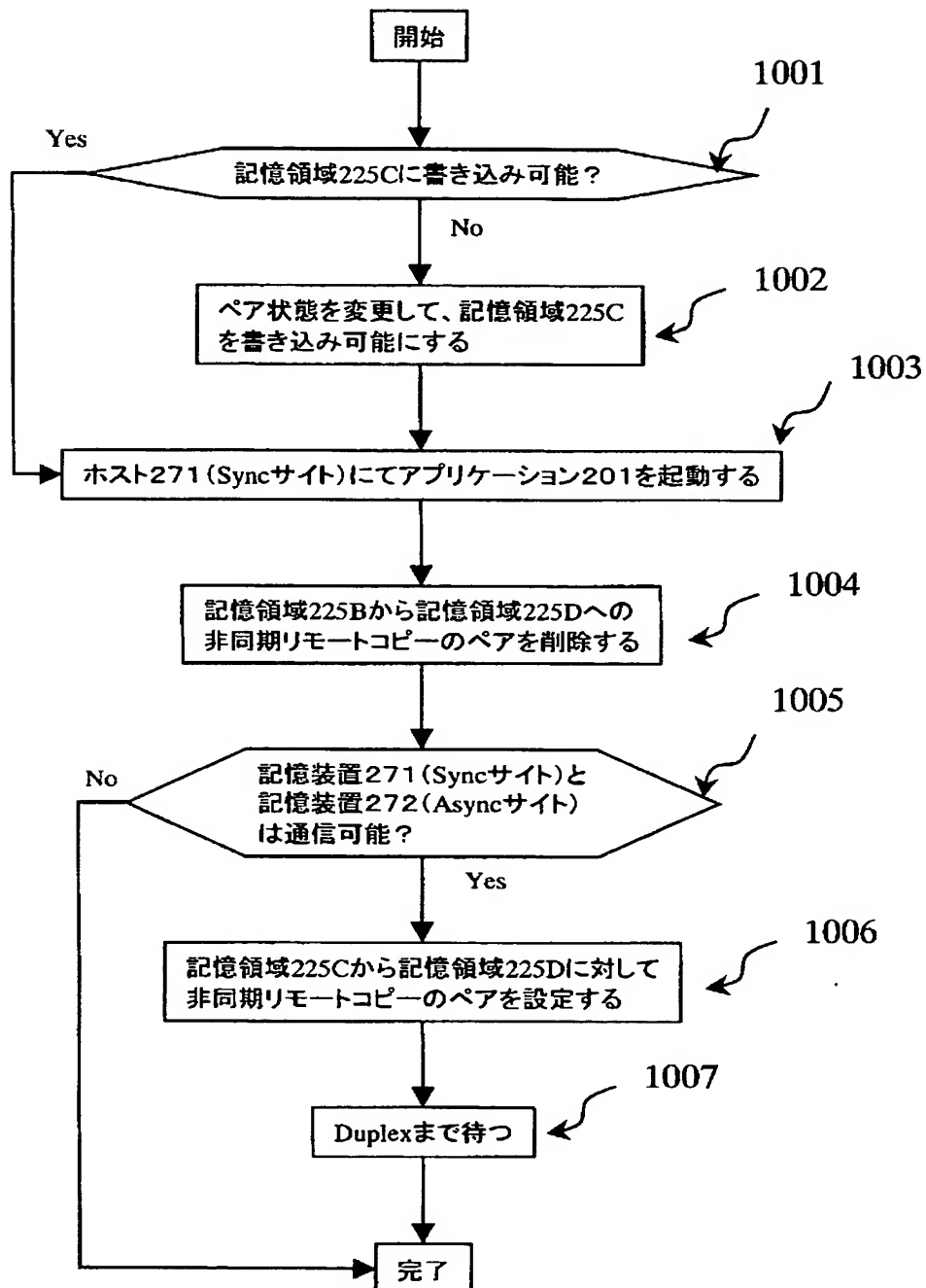
【図9】

図 9



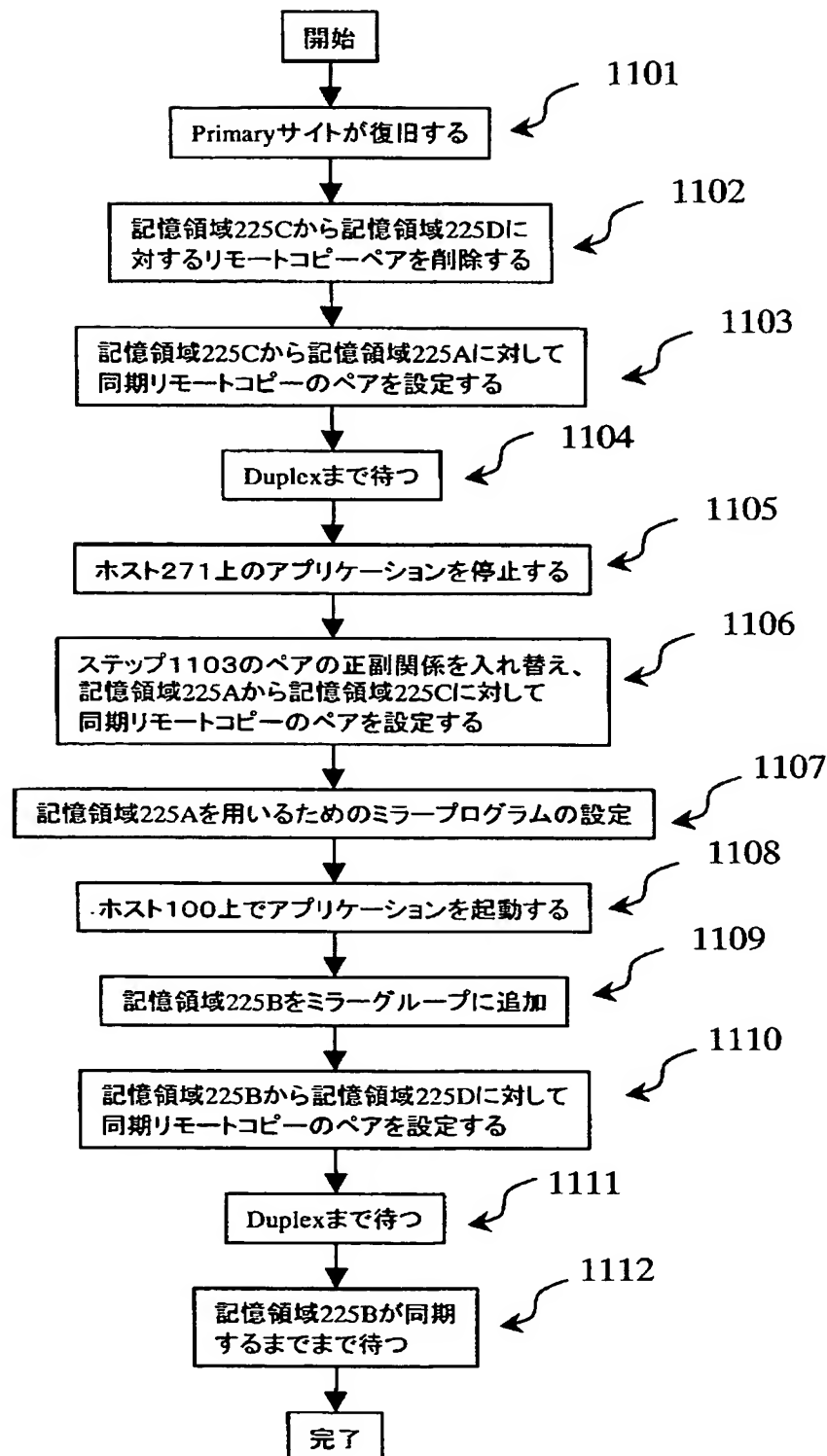
【図10】

図10



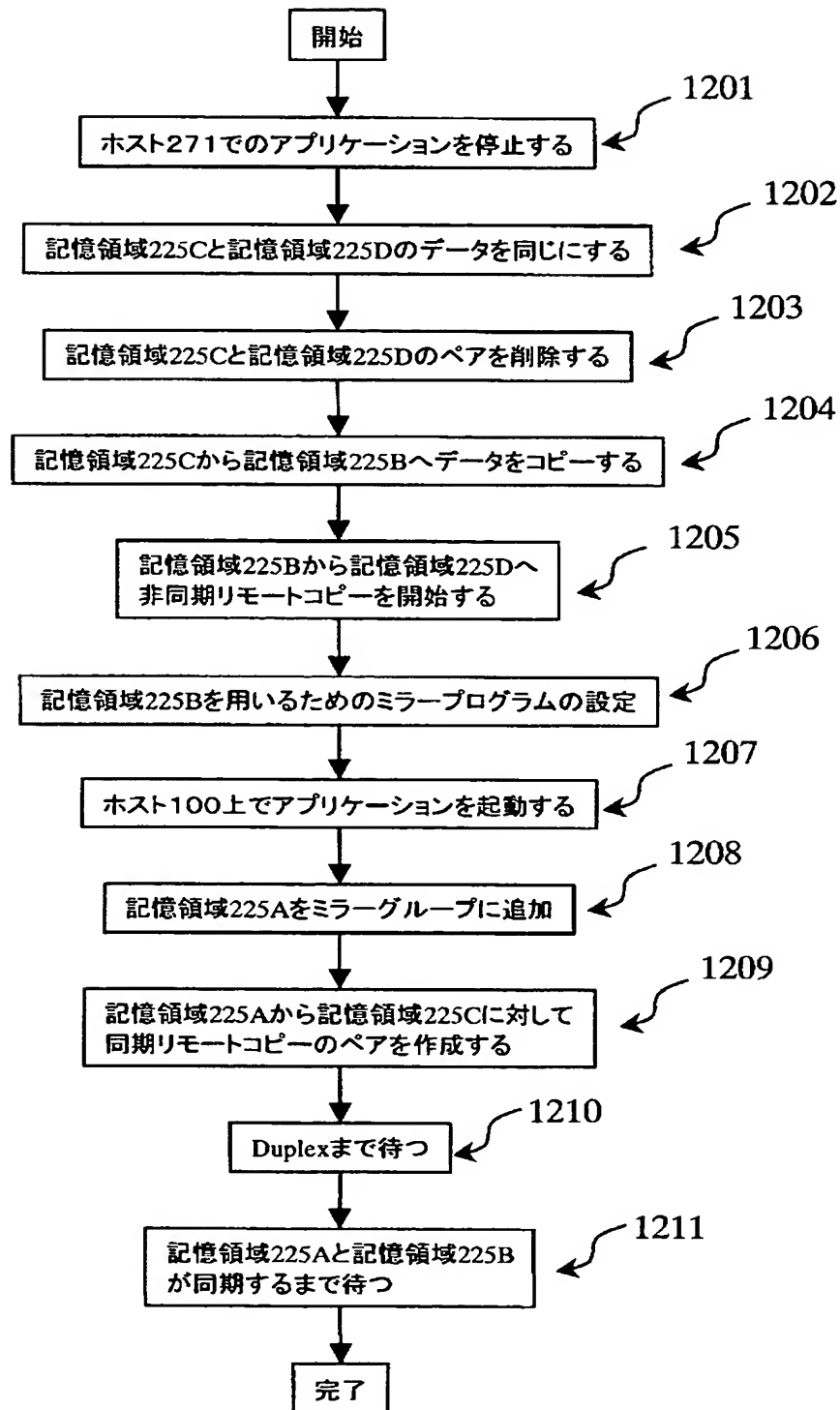
【図 11】

図 11



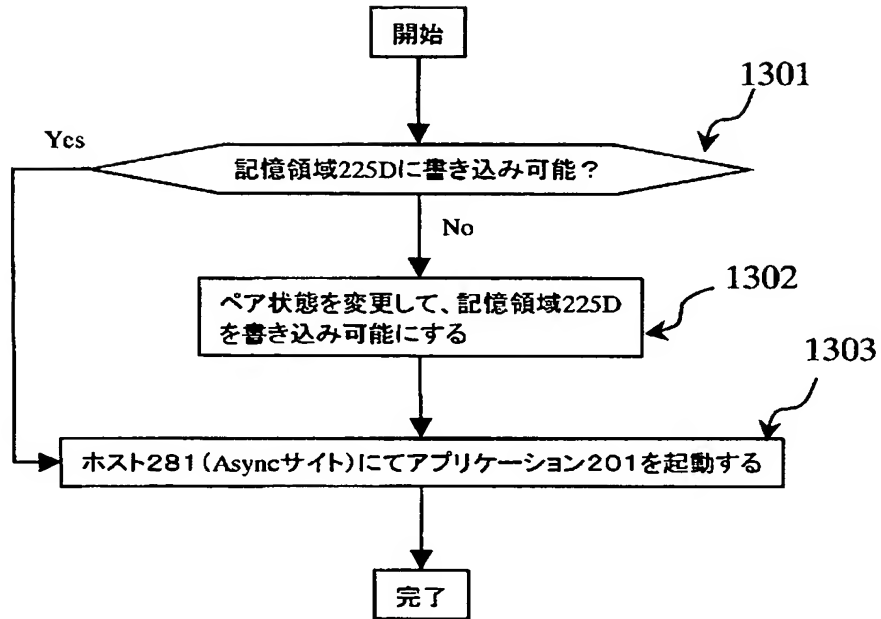
【図 12】

図 12



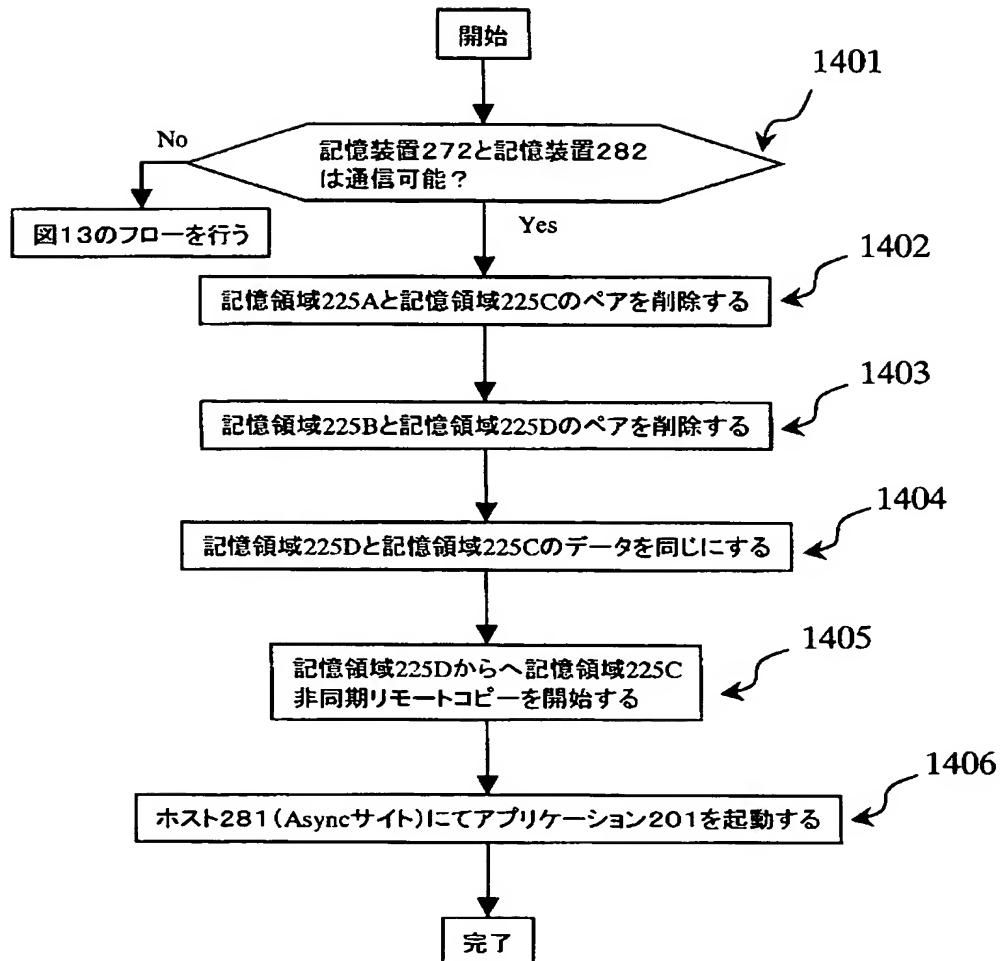
【図 13】

図13



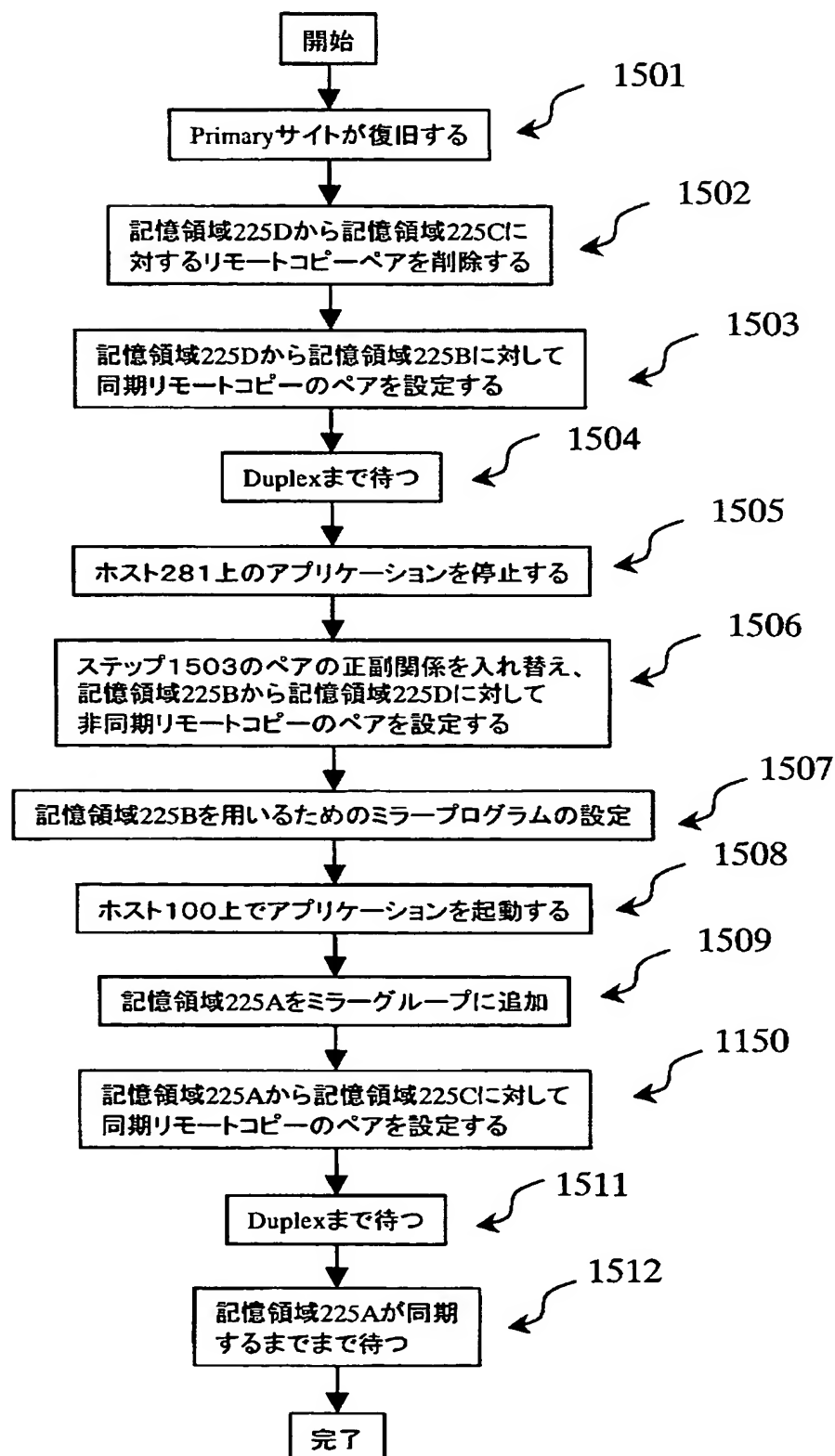
【図 14】

図14



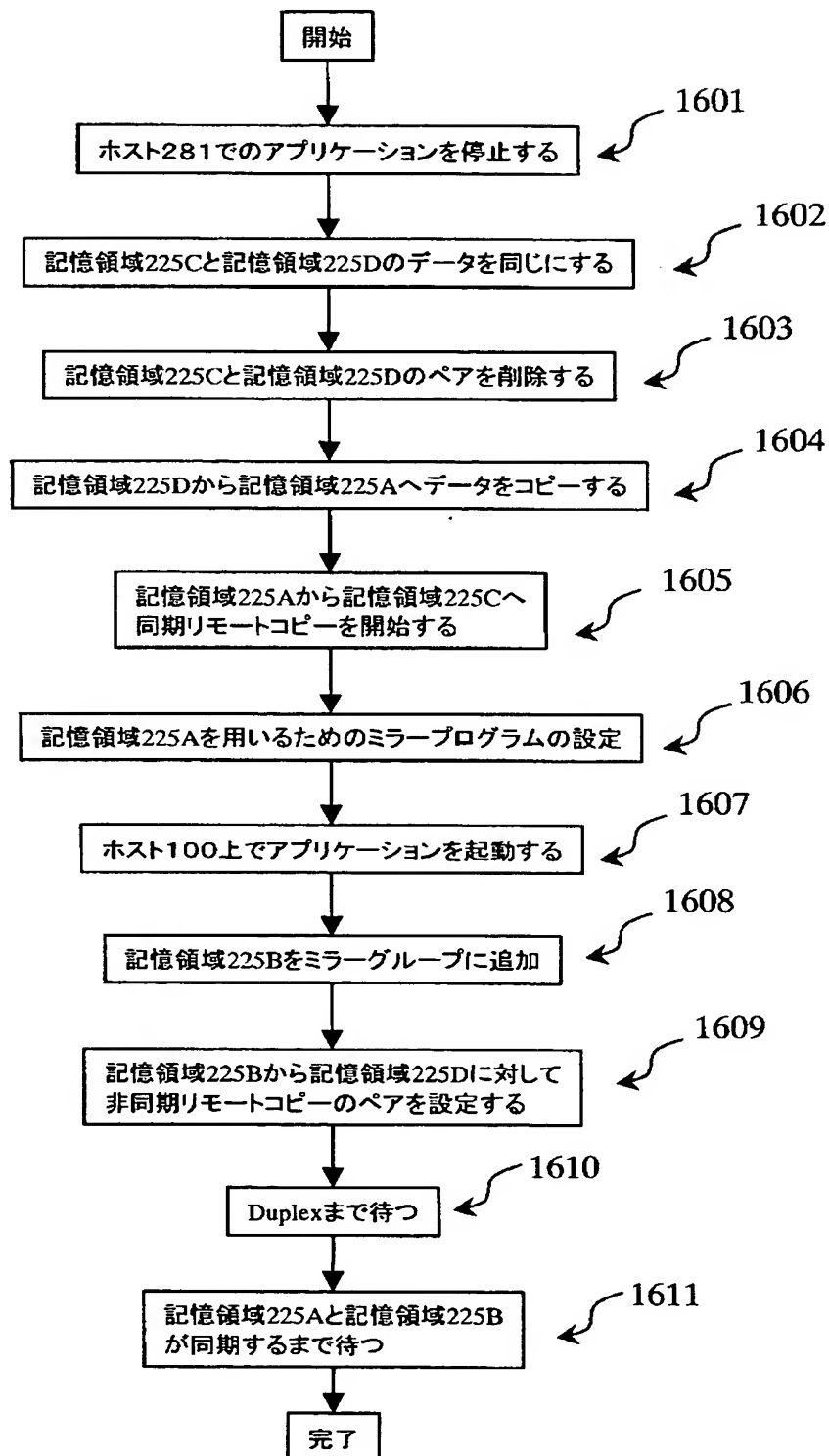
【図15】

図15



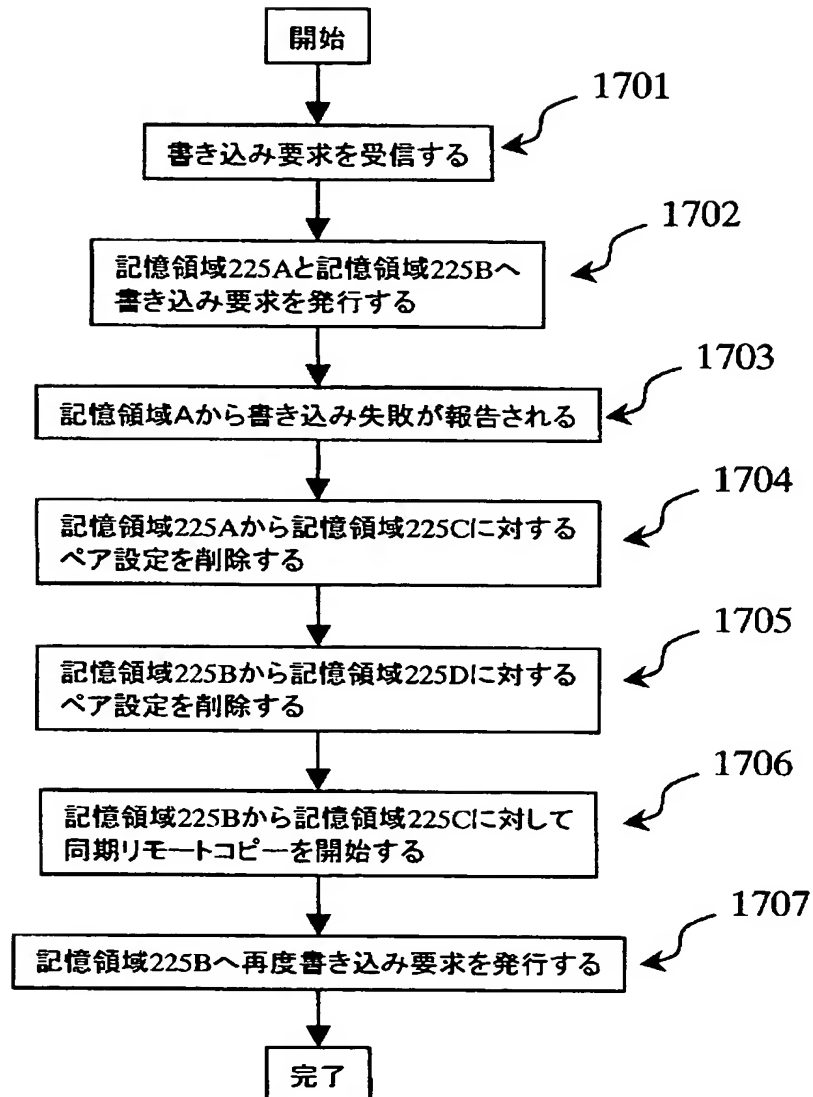
【図16】

図16

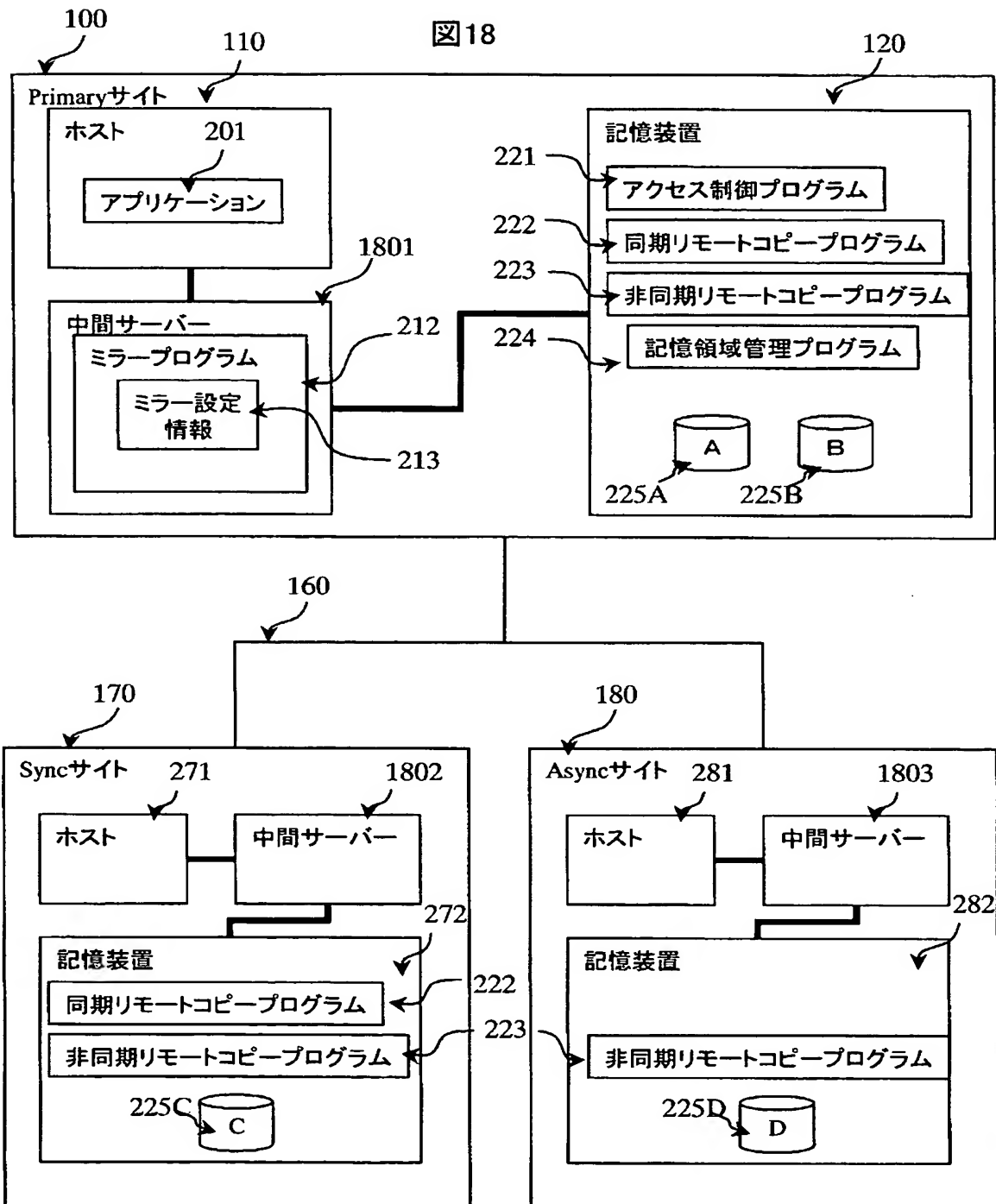


【図 17】

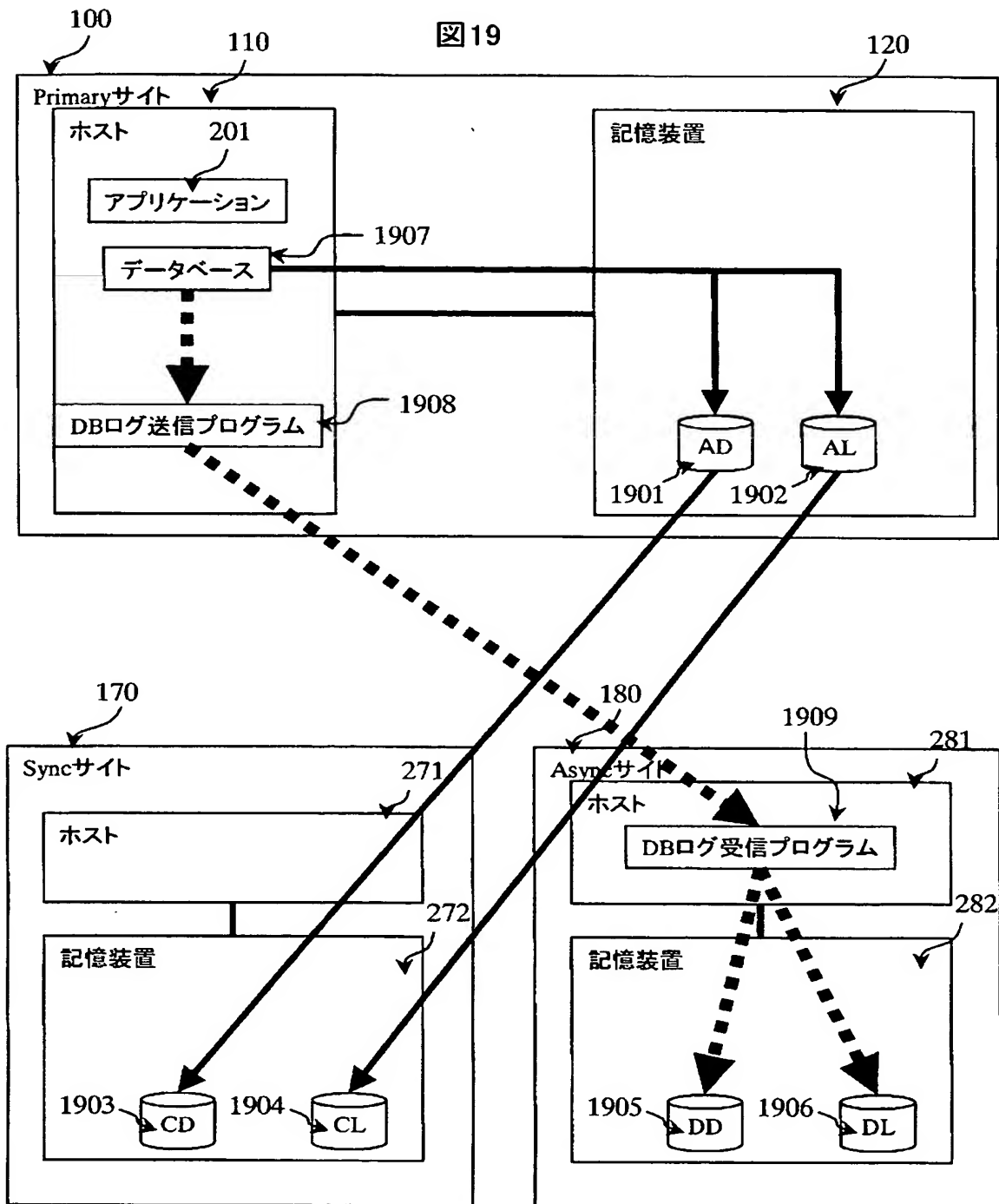
図 17



【図 18】

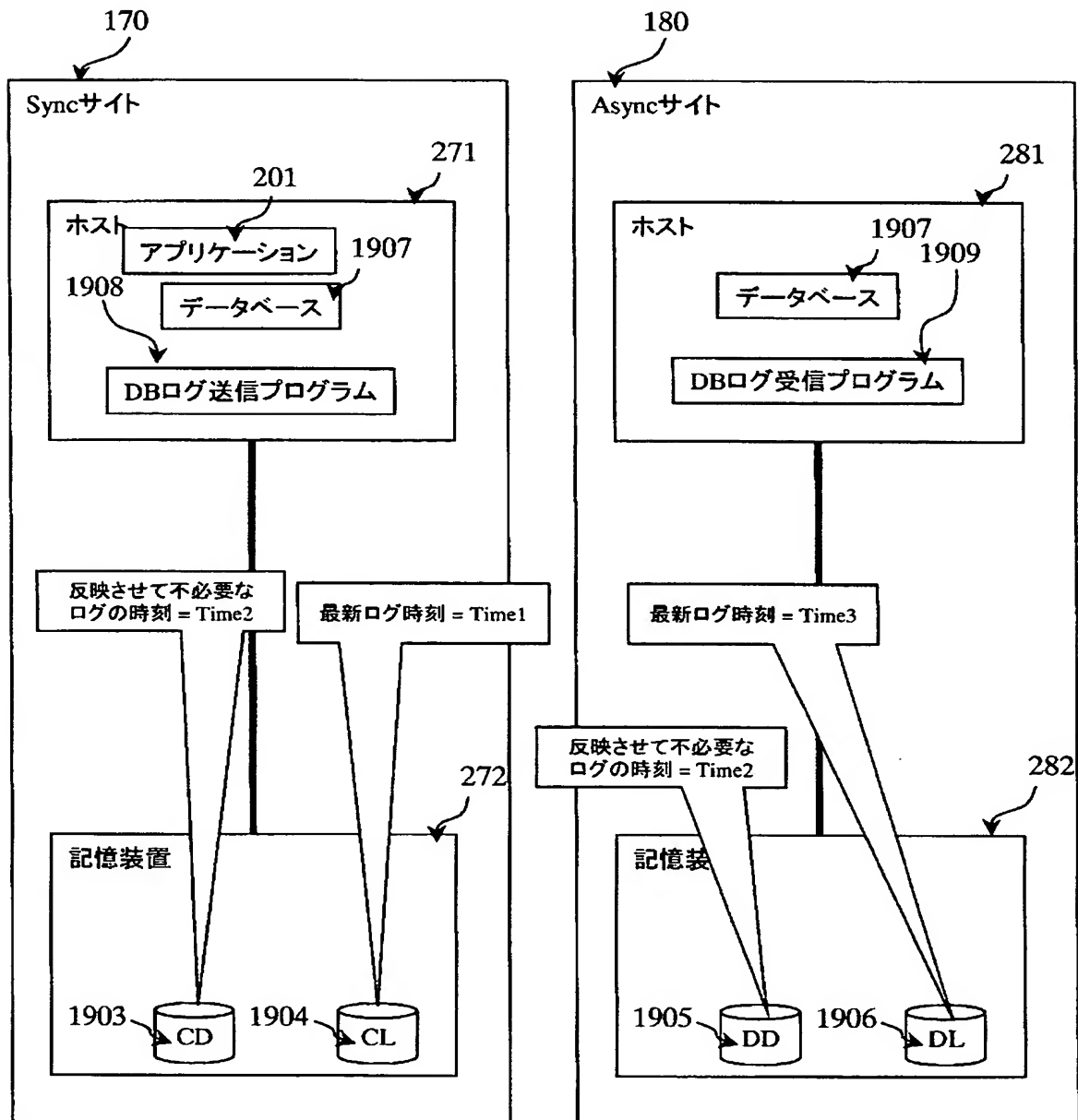


【図 19】

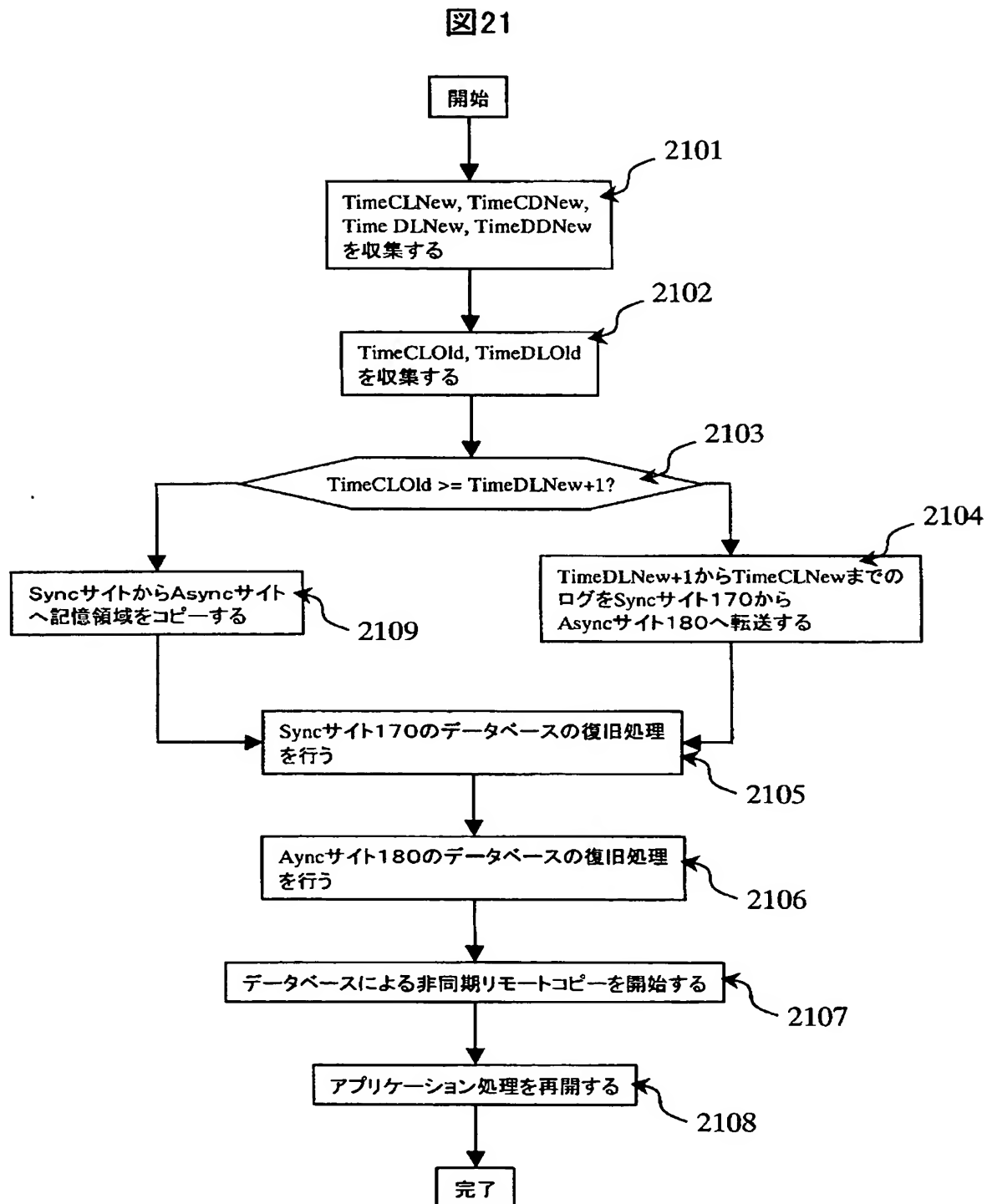


【図 20】

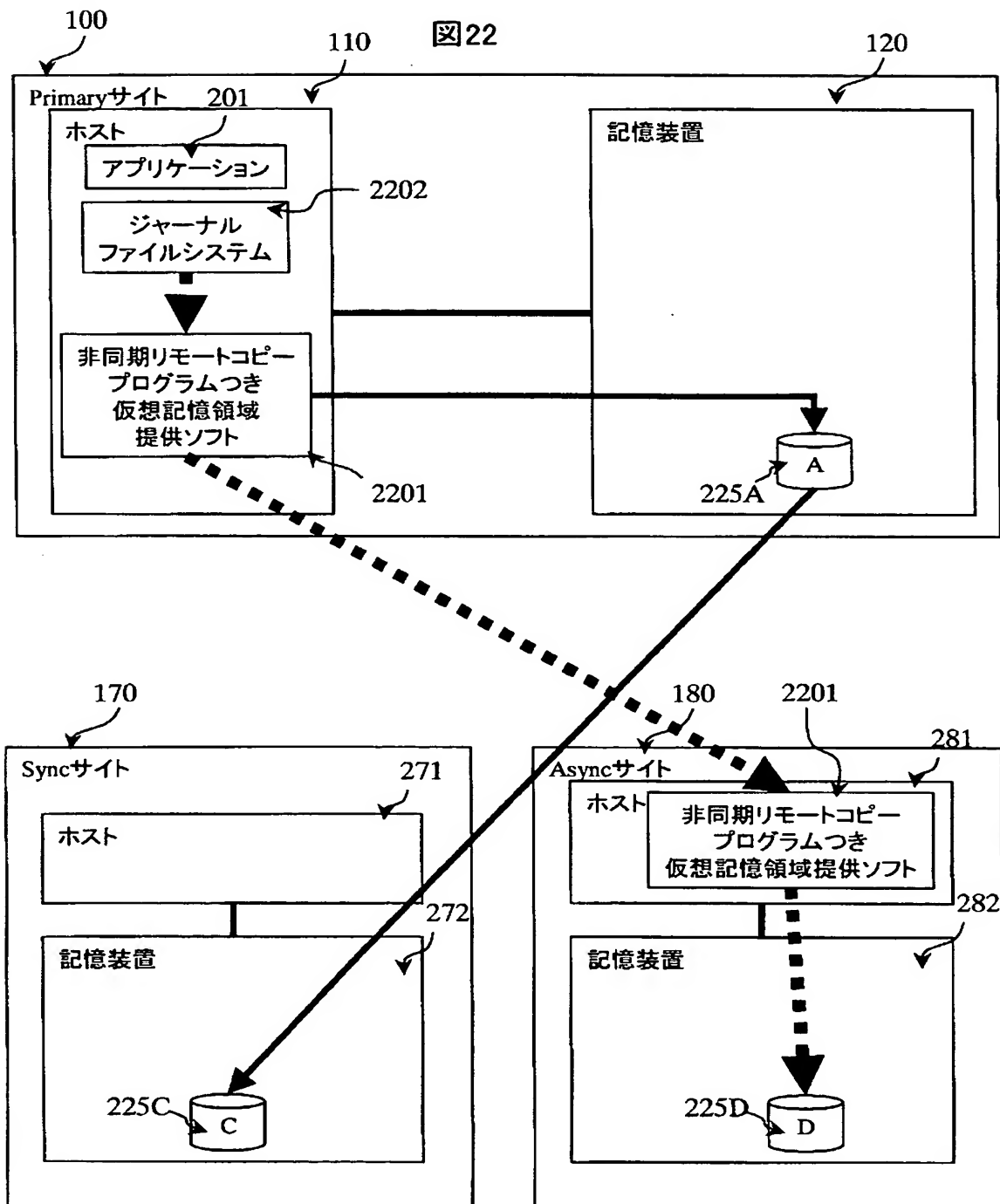
図20



【図 21】



【図 22】



【書類名】 要約書

【要約】

【課題】

第1の記憶装置から遠距離に設置された第3の記憶装置へのデータのコピーが定期的にしか行われないため、第1の記憶装置と第1の記憶装置から近距離に設置された第2の記憶装置が同時に障害になった場合に失う更新データ大きくなる。

【解決手段】

第1の記憶装置に二つの記憶領域を作成し、第1の記憶装置が有する第1の記憶領域から第2の記憶装置が有する記憶領域に対しては同期リモートコピーを行い、第1の記憶装置が有する第2の記憶領域から第3の記憶装置が有する記憶領域に対しては非同期リモートコピーを行う。また、第一の記憶装置にアクセスを行う計算機に第1の記憶装置が有する両記憶領域にミラーリングする。

【選択図】 図2

認定・付加情報

特許出願の番号	特願 2 0 0 3 - 2 0 7 0 0 4
受付番号	5 0 3 0 1 3 2 6 3 0 5
書類名	特許願
担当官	第七担当上席 0 0 9 6
作成日	平成 1 5 年 8 月 1 2 日

< 認定情報・付加情報 >

【提出日】 平成 15 年 8 月 11 日

特願 2 0 0 3 - 2 0 7 0 0 4

出 願 人 履 歷 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所